Published online: 2019-11-10

Patient-Specific Explanations for Predictions of Clinical Outcomes

Mohammadamin Tajgardoon[1] Malarkodi J. Samayamuthu[2] Luca Calzoni[2] Shyam Visweswaran[1,2]

[1] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, Pennsylvania, United States
[2] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Address for correspondence Mohammadamin Tajgardoon, MS, The Offices at Baum, 5607 Baum Boulevard, Suite 523, Pittsburgh, PA 15206, United States (e-mail: mot16@pitt.edu).

ACI Open 2019;3:e88–e97.

Abstract

Background Machine learning models that are used for predicting clinical outcomes can be made more useful by augmenting predictions with simple and reliable patient-specific explanations for each prediction.

Objectives This article evaluates the quality of explanations of predictions using physician reviewers. The predictions are obtained from a machine learning model that is developed to predict dire outcomes (severe complications including death) in patients with community acquired pneumonia (CAP).

Methods Using a dataset of patients diagnosed with CAP, we developed a predictive model to predict dire outcomes. On a set of 40 patients, who were predicted to be either at very high risk or at very low risk of developing a dire outcome, we applied an explanation method to generate patient-specific explanations. Three physician reviewers independently evaluated each explanatory feature in the context of the patient's data and were instructed to disagree with a feature if they did not agree with the magnitude of support, the direction of support (supportive versus contradictory), or both.

Results The model used for generating predictions achieved a F1 score of 0.43 and area under the receiver operating characteristic curve (AUROC) of 0.84 (95% confidence interval [CI]: 0.81–0.87). Interreviewer agreement between two reviewers was strong (Cohen's kappa coefficient = 0.87) and fair to moderate between the third reviewer and others (Cohen's kappa coefficient = 0.49 and 0.33). Agreement rates between reviewers and generated explanations—defined as the proportion of explanatory features with which majority of reviewers agreed—were 0.78 for actual explanations and 0.52 for fabricated explanations, and the difference between the two agreement rates was statistically significant (Chi-square = 19.76, p-value < 0.01).

Conclusion There was good agreement among physician reviewers on patient-specific explanations that were generated to augment predictions of clinical outcomes. Such explanations can be useful in interpreting predictions of clinical outcomes.

Keywords
- predictive model
- patient-specific explanation
- machine learning
- clinical decision support system

received
March 31, 2018
accepted after revision
August 7, 2019

DOI https://doi.org/10.1055/s-0039-1697907.
ISSN 2566-9346.

Background and Significance

Sophisticated predictive models are increasingly being developed using machine learning methods to predict clinical outcomes, such as mortality, morbidity, and adverse events.[1–9] These models, in most cases, are viewed as black boxes that produce a prediction for an outcome from the features of a patient case.[†] However, for such models to be practically useful in clinical care, it is critical to provide clear and reliable individual-specific explanations for each prediction.[10] While a prediction provides an estimate of the likely outcome in the future, an explanation provides insight into features that may be useful in clinical decision-making. Moreover, explanations will enable physicians to engender trust in the predictions, interpret them in the clinical context, and help make optimal clinical decisions.[11] In the clinical context, features that are supportive of a prediction provide potentially actionable aspects that may change the predicted outcome.[12,13]

In the context of predictive models, a subtle but important distinction exists between model explanation and prediction explanation. Model explanation provides an interpretation of the model to the user in terms of structure and parameters, and is useful in the context of making discoveries.[12,14] Some predictive models, such as decision trees, linear regression, and rule-based models, are more easily interpretable, though often such models have poorer predictive performance than more abstract models, such as random forests, support vector machines, and neural networks.[12,14] In contrast to model explanation, prediction explanation provides an interpretation of the prediction for an individual to whom a model is applied, and will potentially be different from individual to individual.[15,16] Useful prediction explanations possess two properties. First, an explanation uses concepts that are understandable to the user, such as clinical variables that are not modified or transformed. Second, the explanation is parsimonious, so that it is readily and rapidly grasped by the user. Prediction explanations may be based on the structure and parameters of the model that yielded the prediction (hence, model dependent) or may be based on an independent method that is applied after the predictive model has been developed (hence, model independent).[14,17]

Novel methods have been developed for prediction explanations and such methods have been applied in biomedicine and other domains. ►Table 1 provides a summary of studies that have developed methods for prediction explanations, with a brief description of each explanation method.

Only a small number of the methods that are listed in ►Table 1 have been applied to predicting clinical outcomes. For example, Luo applied their method to type-2 diabetes risk prediction[18], Štrumbelj et al developed and applied their method to breast cancer recurrence predictions,[19] and Reggia and Perricone developed explanations for predictions of the type of stroke.[11] More widespread application of these methods to clinical predictions can provide evidence of applicability and utility of these methods to clinical users.

In this study, we apply and evaluate a recently developed prediction explanation method called Local Interpretable Model-Agnostic Explanations (LIME)[15] for clinical predictions. The developers of LIME demonstrated that human evaluators found explanations generated by LIME to be more reasonable when compared with the explanations generated by alternative methods. To our knowledge, LIME has not been extensively evaluated in the context of clinical predictive models.

Objectives

Our goal was to evaluate patient-specific explanations for clinical predictions. The aims of our study were to (1) Develop machine learning models to predict dire outcomes (severe complications including death) from readily available clinical data in patients who present with community acquired pneumonia (CAP), followed by application of a model-independent prediction explanation method to generate patient-specific explanations; and (2) Evaluate the agreement among physicians for explanations generated for CAP patients who were predicted to be either at very high risk or at very low risk of developing a dire outcome.

Methods

In this section, we briefly describe the pneumonia dataset that we used in the experiments, the methods for development and evaluation of predictive models, the generation of patient-specific explanations, and the measures we used to evaluate agreement among physician reviewers for the explanations. The implementation of the methods is publicly available at: https://github.com/Amin-Tajgardoon/explanation-project.

Description of Dataset

The pneumonia data were collected by the Pneumonia Patient Outcomes Research Team (PORT)[20] during October 1991 to March 1994 at five hospitals in three geographical locations including Pittsburgh, Boston, and Halifax, Nova Scotia. The PORT data from Pittsburgh that we used in the experiments had 2,287 patients diagnosed with CAP who were either hospitalized or seen in ambulatory care. A variety of clinical data were collected at the time of presentation and several outcomes at 30 days were assessed. A key goal of the PORT project was to develop accurate criteria for prognosis of patients with pneumonia that could provide guidance on which patients should be hospitalized and which patients might be safely treated at home.

The PORT dataset contains more than 150 variables including demographic information history and physical examination information, laboratory results, and chest X-ray findings. From the 150 variables, we selected 41 clinical variables that are typically available in the emergency

[†] We distinguish between a variable and a feature. A variable describes an aspect of an individual. A feature is the specification of a variable and its value. For example, "fever" is a variable and "fever = yes" is a feature.

Table 1 Studies that describe methods for prediction explanation

Author (year)	Title	Description of method
Lundberg and Lee (2017)	A unified approach to interpreting model predictions[30]	Presents a unified framework for six prediction explanation methods. Also, proposes a new explanation method that outperforms prior methods in terms of computational complexity and reliability.
Krause et al (2016)	Interacting with predictions: visual inspection of black-box machine learning models[31]	Describes an interactive environment that enables the user to inspect a model's prediction by tweaking feature values and observing the effect on the model's behavior.
Luo (2016)	Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction[18]	Develops a rule-based model to explain the decision made by the prediction model.
Ribeiro et al (2016)	"Why should I trust you?": Explaining the predictions of any classifier[15]	Proposes a post-hoc explanation method that generates data samples that are similar to the predicted sample, labels the samples by the predictive model, and fits a local linear model to the samples. Uses the weights in the local model to identify the influential features.
Baehrens et al (2009)	How to explain individual classification decisions[32]	Proposes a prediction explanation method that uses the gradient vector of the predictive model at the point of the predicted sample for measuring feature importance.
Sikonja and Kononenko (2008)	Explaining classifications for individual instances[33]	Explains a sample by assigning an importance factor to each sample's feature. The importance factor of a feature is defined as the change in the model's prediction on removal of the feature from the sample.
Štrumbelj and Kononenko (2008)	Toward a model independent method for explaining classification for individual instances[34]	Describes a model-independent explanation method for probabilistic classifiers. Calculates an importance weight for each feature by measuring the change in the class probability on removal of the feature from the conditional probability of the class given the sample features.
Lemaire et al (2008)	Contact personalization using a score understanding method[35]	Computes the influence of a feature by measuring the effect of changing the feature's value on the model's prediction.
Poulin et al (2006)	Visual explanation of evidence in additive classifiers[36]	Describes a framework to visualize each feature's contribution to a prediction. Provides the capability to analyze the effect of changing feature values on a classifier's decision. The method is applicable to additive models such as naïve Bayes, and support vector machines.
Szafron et al (2003)	Explaining naïve Bayes classifications[37]	Provides a graphical explanation framework for naïve Bayes predictions. For a sample, the framework visualizes each feature's contribution to the decision made by the classifier.
Reggia and Perricone (1985)	Answer justification in medical decision support systems based on Bayesian classification[11]	Proposes an explanation method for Bayesian classifiers by using prior and likelihood values to determine important features responsible for the posterior probability of the outcome.

department at the time the decision whether to admit or not is made. Of the 41 variables, 17 are discrete and the remaining 24 are continuous. The 24 continuous variables were discretized based on thresholds provided by clinical experts on the PORT project.[20] A list of the 41 variables with descriptions is provided in ►Table 2.

The outcome variable we used as the target variable is called dire outcome and is binary. A patient was considered to have had a dire outcome if any of the following events occurred: (1) death within 30 days of presentation; (2) an intensive care unit admission for respiratory failure, respiratory or cardiac arrest, or shock; or (3) one or more specific,

Table 2 list of variables in the pneumonia PORT study that were used in the present study

Domain	Variable	Description
Demographics	Age	(Discrete) [1–6] Range was [18–105]
	Sex	Female/male
	Race	White/non-white
	Ethnicity	Hispanic/non-Hispanic
	Smoking status	Yes/no
Past history	Number of prior episodes of pneumonia	[0–2]
Comorbidities	Congestive heart failure	Yes/no
	Cerebrovascular disease	Yes/no
	Liver disease	Yes/no
	Cancer	Yes/no
Symptoms	Cough	Yes/no
	Fever	Yes/no
	Sweating	Yes/no
	Headache	Yes/no
Physical exam	Confusion	Yes/no
	Lungs status	Clear/congested
Vitals	HR (heart rate)	(Discrete) [1–3]
	BP systolic (systolic blood pressure)	(Discrete) [1–3]
	BP diastolic (diastolic blood pressure)	(Discrete) [1–3]
	RR (respiratory rate)	(Discrete) [1–3]
	Temp (temperature)	(Discrete) [1–5]
Laboratory results	WBC (white blood cell count)	(Discrete) [1–5]
	Hgb (hemoglobin)	(Discrete) [1–3]
	Hct (hematocrit)	(Discrete) [1–4]
	Plt (Platelet count)	(Discrete) [1–4]
	Na (sodium)	(Discrete) [1–4]
	K (potassium)	(Discrete) [1–3]
	HCO_3 (bicarbonate)	(Discrete) [1–3]
	BUN (blood urea nitrogen)	(Discrete) [1–4]
	Cr (creatinine)	(Discrete) [1–3]
	Glu (glucose)	(Discrete) [1–4]
	Tot Bili (total bilirubin)	(Discrete) [1–3]
	SGOT/AST (aspartate aminotransferase)	(Discrete) [1–3]
	Alk Phos (alkaline phosphatase)	(Discrete) [1–3]
	LDH (lactate dehydrogenase)	(Discrete) [1–3]
ABG (arterial blood gas)	pH	(Discrete) [1–4]
	pCO_2	(Discrete) [1–4]
	pO_2	(Discrete) [1–4]
	O_2 saturation	(Discrete) [0–1]
X-ray	Infiltrate	Yes/no
	Pleural effusion	Yes/no
Outcome	Dire outcome	Yes/no

Note: Continuous variables were discretized based on clinical judgment of pneumonia experts in the pneumonia PORT project.[20] The label "(Discrete)" in the description indicates that a variable is a discretized version of a continuous variable.

severe complications, such as myocardial infarction, pulmonary embolism, stroke, etc.[21] About 11.4% (261) patients had a dire outcome in the PORT dataset.

Training and test sets: the data consisting of 2,287 cases was divided into a training dataset of 1,601 cases (70%) and a test dataset of 686 cases (30%) by using stratified random-sampling such that both sets had approximately the same proportion of cases with dire outcomes as the full dataset (11.4% [182/1,601] and 11.5% [79/686] of patients had a dire outcome in the training and test sets, respectively). Missing data were imputed using an iterated *k*-nearest neighbor method,[22] and continuous variables were discretized based on clinical judgment of pneumonia experts in the pneumonia PORT project.

Development of Predictive Models

We applied several machine learning methods to the training set to develop predictive models, and we applied the best-performing model to the test set to generate predictions.

Machine learning methods: the machine learning methods that we used for developing predictive models are logistic regression with regularization (LR), random forest (RF), support vector machine (SVM), and naïve Bayes (NB). We selected these methods as representative of the machine-leaning methods that are typically used for developing predictive models in biomedicine. We used the implementations of these methods that are available in the scikit-learn package.[23]

We tuned the hyper-parameters using 10-fold cross validation on the training set. The hyper-parameters that we configured included the regularization coefficient ([0.1, 1, 10]) for the LR and SVM models, number of trees ([100, 500, 1,000, 3,000]) for the RF model, and the Laplace smoothing parameter ([0, 0.1, 1, 10, 100]) for the NB model.

Evaluation of model performance: we evaluated the predictive models on the training set using 10-fold cross validation. The metrics we used included F1 score, area under the receiver operating characteristic curve (AUROC), positive predictive value (PPV), sensitivity, and specificity. The F1 score is the harmonic mean of PPV and sensitivity and ranges between 0 and 1.[24] A high F1 score indicates that both PPV and sensitivity are high. We selected the machine learning method with the highest F1 score and reapplied it to the full-training set to derive the final model. We applied the final model to predict the outcomes for cases in the test set.

Generation of Patient-Specific Explanations

We used LIME to generate explanations for a selected set of 40 cases in the test set. A description on the selection of the 40 cases is provided in the next section. LIME is a model-independent explanation method that provides an explanation for a predicted case by learning an interpretable model from data in the neighborhood of the case (such as a local linear model with a small number of nonzero coefficients). More specifically, LIME provides for each patient feature the magnitude and the direction of support for the predicted outcome (see ►Fig. 1). The magnitude of support is the weight of an explanatory feature, and the direction of support is the sign of the weight, as estimated in LIME's local regression model. We limited the explanations to the top six features with the highest magnitudes, as we found that, on average, the magnitude of five to seven features accounted for most of the total magnitude. We call the patient features that were included in the explanation as explanatory features.

Evaluation of Explanations

Three physicians independently evaluated explanations for 40 patient cases that were selected from the test set. We selected cases for which the model correctly predicted the outcome with high confidence (i.e., a patient was predicted to have developed a dire outcome with probability > 0.8 or with probability < 0.1). Of the 40 cases, 20 patients developed a dire outcome and 20 patients did not. Note that patients with and without a dire outcome are expected to have mostly the same predictors; however, the values of

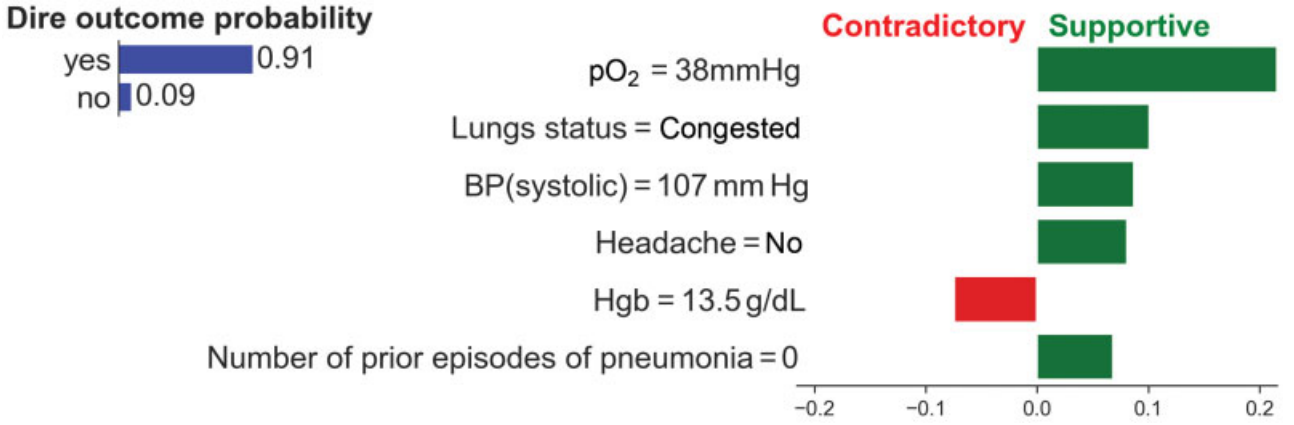


Fig. 1 Example explanation obtained from LIME for a patient who was predicted to have a very high probability of dire outcome by a logistic regression model. The bar plot at the top left shows the predicted probability distribution for dire outcome. The bar plot on the right shows the explanation for the prediction. The explanation is limited to six top-ranked features by magnitude. The magnitude on the horizontal axis represents the weight of a feature in the LIME's local regression model. Green bars represent the magnitude of predictors that support the predicted outcome, while red bars represent the magnitude of contradictory features. LIME, local interpretable model-agnostic explanations.

those predictors are likely to be different. For example, abnormal values in respiratory rate, arterial blood gases, and lung status are likely to be predictor features in a patient with a dire outcome, whereas normal values in respiratory rate, arterial blood gases, and lung status are likely to be predictor features in a patient without a dire outcome.

For each patient case, we provided the reviewers with a description that included clinical findings and if a dire outcome occurred or not, and the predicted probability of the dire outcome occurring along with the explanation for the prediction (see ►Fig. 2). Each reviewer assessed all 40 cases and the corresponding explanations, and specified if she agreed or disagreed with each explanatory feature. The reviewer was instructed to disagree with an explanatory feature if she did not agree with the magnitude, the direction (supportive vs. contradictory), or both.

To preclude reviewers from agreeing readily with explanations without careful assessment, we fabricated explanations in 10 of the 40 cases. To generate a fabricated explanation, we replaced the labels (feature name and its value) of six top-ranked features with the labels of six bottom-ranked features, without modifying the magnitude or the direction of support. The reviewers were informed that some of the cases contained fabricated explanations but not which ones. ►Table 3 shows the stratification of cases according to the type of explanation (actual vs. fabricated) and by outcome (had a dire outcome vs. did not have a dire outcome) that we used for evaluation.

Patient description

Demographics:
Age = 63 years, Sex = Female, Race = White , Ethnicity = Non-hispanic, Smoking status = Non-smoker

Past history:
Number of prior episodes of pneumonia = 0

Comorbidities:
Congestive heart failure = Yes, Cerebrovascular disease = No, Liver disease = No, Cancer = No

Symptoms:
Cough = Yes, Fever = Yes, Sweating = Yes, Headache = No

Physical exam:
Confusion = No, Lungs status = Congested

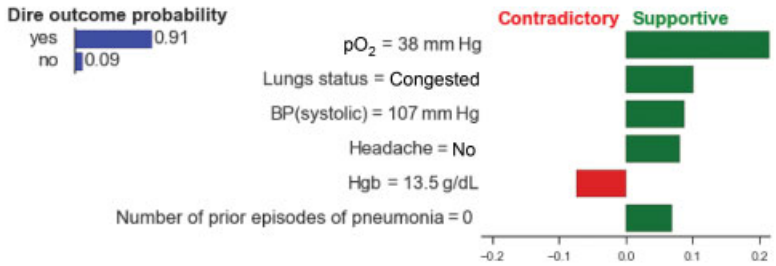
Vitals:
HR = 118/min BP = 107/42 mmHg RR = 32/min Temp = 37.86° C

Laboratory results:
WBC = 5.7x10^3/uL Hgb = 13.5 g/dL Hct = 41.0% Plt = 156x10^3/uL
Na = 128 mEq/L K = 5.3 mEq/L HCO3 = 34 mEq/L BUN = 27 mg/dL Cr = 1.3 mg/dL Glu = 727 mg/dL
Tot bili = 0.6 mg/dL SGOT/AST = 45 U/L Alk Phos = 108 U/L LDH = 518 U/L
ABG: pH = 7.45 pCO2 = 47 mmHg pO_2 = 38 mmHg O2 saturation = 72% (pulse ox)

X-ray:
Infiltrate = Yes, Pleural effusion = No

Outcome:
Dire outcome = Yes

Explanation



In the table below in the last column, please state your agreement/disagreement with each feature. You may indicate agreement with + and disagreement with –.

Feature	Agree/Disagree
pO2 = 38 mmHg	
Lungs status = Congested	
BP(systolic) = 107 mmHg	
Headache = No	
Hgb = 13.5 g/dL	
Number of prior episodes of pneumonia = 0	

Fig. 2 An example patient case that gives a description of the patient, followed by an explanation and the questions that were asked of reviewers.

Table 3 Cases used for evaluation, stratified by type of explanations and outcomes

Type of explanations and outcomes	Number of cases
Cases with actual explanations	
where patients had a dire outcome	15
where patients did not have a dire outcome	15
Cases with fabricated explanations	
where patients had a dire outcome	5
where patients did not have a dire outcome	5
Total	40

We analyzed the assessments of the reviewers with several measures as follows: (1) We measured agreement between pairs of reviewers with Cohen's kappa coefficient[25] and across all reviewers concurrently with Fleiss' kappa statistic.[26] Cohen's kappa coefficient measures the degree of agreement between two reviewers on a set of samples, whereas Fleiss' kappa statistic can assess more than two reviewers simultaneously. (2) For a given set of cases, we calculated an agreement rate as the proportion of explanatory features with which majority of reviewers agreed. For example, for a set of 10 cases where each case had an explanation with six features, the denominator of the agreement rate is $10 \times 6 = 60$ features and the numerator is the number of features with which majority of reviewers agreed. Agreement rates were calculated separately for cases with actual and fabricated explanations, and for cases where the patients had a dire outcome and did not have a dire outcome. (3) To statistically test for difference between two agreement rates that are derived from two sets of cases (e.g., actual vs. fabricated explanations, dire outcome vs. no dire outcome), we used the Chi-square test of independence.[27]

Results

We report the performance of the machine learning methods, briefly describe the prediction explanations, and provide the reviewers' agreement scores.

Performance of Predictive Models

►Table 4 shows the performance of five machine learning methods on the training set, as measured by F1 score, AUROC, PPV, sensitivity, and specificity. The two logistic regression models, LR-L1 and LR-L2, were trained with L1 and L2 regularization penalties, respectively. The LR-L1, LR-L2, NB, and SVM models have similar F1 scores, whereas RF has a lower F1 score despite having a similar AUROC to other models. The LR-L1 and LR-L2 models had similar performance; however, we chose the LR-L1 model as the best-performing model because it shrinks some of the regression coefficients to zero and provides a sparse solution.

Description of Explanations

We applied the LR-L1 model to all cases in the test set and selected 40 cases based on criteria described in Section Methods, "Evaluation of Explanations." We used LIME to generate explanations for the selected cases. ►Tables 5 and 6 show the explanatory variables and their count of appearance in the actual and fabricated explanations respectively.

Evaluation of Explanations

Agreement among reviewers: ►Table 7 shows the agreement scores between pairs of reviewers and across all three reviewers. For both actual and fabricated explanations, Cohen's kappa coefficients indicate strong agreement between reviewers 1 and 2, and fair to moderate agreement between reviewer 3 and the other two reviewers (according to the agreement levels proposed by McHugh[28]). The Fleiss' kappa statistic shows moderate agreement across all reviewers when considering all explanatory features. Much of the disagreement between reviewer 3 and the others was due to differing opinions on headache as an explanatory feature. After excluding headache from the analysis, Cohen's kappa coefficient for all explanatory features for reviewers 1 and 3 increased from 0.49 to 0.76, and the corresponding Cohen's kappa coefficient for reviewers 2 and 3 increased from 0.33 to 0.58.

Agreement with LIME-generated explanations: ►Table 8 shows agreement rates for explanations as the proportion of explanatory features with which majority of reviewers agreed. The agreement rate was 0.78 (141/180) for actual explanations and 0.52 (31/60) for fabricated explanations;

Table 4 Performance of five machine learning methods on the training set using 10-fold cross validation

Model	F1 score	AUROC	PPV	Sensitivity	Specificity
LR-L1	0.43 (± 0.02)	0.84 (± 0.03)	0.31 (± 0.02)	0.69 (± 0.02)	0.81 (± 0.02)
LR-L2	0.43 (± 0.02)	0.84 (± 0.03)	0.32 (± 0.02)	0.69 (± 0.02)	0.81 (± 0.02)
NB	0.42 (± 0.02)	0.84 (± 0.03)	0.30 (± 0.02)	0.76 (± 0.02)	0.76 (± 0.02)
SVM	0.42 (± 0.02)	0.84 (± 0.03)	0.29 (± 0.02)	0.74 (± 0.02)	0.77 (± 0.02)
RF	0.23 (± 0.02)	0.85 (± 0.03)	0.52 (± 0.02)	0.16 (± 0.02)	0.98 (± 0.01)

Abbreviations: AUROC, area under the receiver operating characteristic curve; CI, confidence interval; LR-L1, LASSO logistic regression; LR-L2, ridge logistic regression; NB, naïve Bayes; PPV, positive predictive value; RF, random forest; SVM, support vector machine.
Note: The models are sorted in descending order of their F1 scores. The 95% CI for AUROCs were calculated using the Delong's method,[38,39] and the 95% CI for the other measures were calculated using the Wilson's score interval.[40]

Table 5 Variables and their count of appearance in the 30 actual explanations

Variable	Count
Lungs status	30
Headache	30
pO_2 (arterial blood gas)	23
RR (respiratory rate)	21
Prior episodes of pneumonia	18
Hgb (hemoglobin)	18
Glu (glucose)	17
BP systolic	16
Age	5
Sweating	1
Confusion	1

Table 6 Variables and their count of appearance in the 10 fabricated explanations

Variable	Count
Sex	10
Race	7
Cr (creatinine)	7
K (potassium)	6
HR (heart rate)	6
Plt (platelet count)	5
pCO_2 (arterial blood gas)	4
WBC (white blood cell count)	4
BP (diastolic)	4
Ethnicity	3
Hct (hematocrit)	2
Liver disease	1
Infiltrate	1

the difference between the two agreement rates was statistically significant (Chi-square = 19.76, p-value < 0.01). For actual explanations, agreement rates were 0.81 (73/90) for cases where the patients had a dire outcomes and 0.76 (68/90) for cases where the patients did not have a dire outcome; the difference between the two agreement rates was not statistically significant (Chi-square = 0.55, p-value = 0.53).

When headache was excluded from the analysis, the agreement rate increased from 0.78 to 0.93 for actual explanations. The agreement rate for fabricated explanations did not change from 0.52 because headache did not occur in fabricated explanations.

Discussion

Computerized clinical decision-supporting systems that utilize predictive models for predicting clinical outcomes can be enhanced with explanations for predictions. Such explanations provide context for the predictions and guide physicians in better understanding supportive and contradictory evidence for the predictions. In this paper, we presented a method to augment clinical outcome predictions—obtained from a predictive model—with simple patient-specific explanations for each prediction. The method uses LIME that generates a patient-specific linear model which provides a weight for each feature. The weight provides insight about the relevance of each feature in terms of magnitude and direction of its contribution to a prediction. LIME has been shown to produce explanations that users find to be useful and trustworthy in general prediction problems.[15]

In this study, we developed and evaluated several machine learning methods and chose a logistic regression model since it had the best performance. In this scenario, the model could be used directly to provide explanations—the weight of a feature for an explanation can be computed by multiplying the feature level by the corresponding odds ratio. However, in general, as the size and dimensionality of the data increase, more complex, and less interpretable models, like deep neural networks, are likely to perform better and the use of a model-independent explanation method like LIME becomes necessary.

Using LIME, we generated explanations for 40 cases and evaluated the explanations with three physician reviewers. The reviewers agreed with 78% of LIME-generated explanatory features for actual explanations and agreed with only 52% of explanatory features for fabricated explanations. This result provides evidence that the reviewers are able to distinguish between valid and invalid explanations. The results also indicate that agreement on cases where the patients had a dire outcome is not statistically significantly different from agreement on cases where the patients did not have a dire outcome.

Headache was a feature that was provided as an explanatory feature in most of the cases where the patients experienced a dire outcome. Two of the reviewers deemed headache to be mildly supportive, whereas the third reviewer did not consider headache to be a supportive feature. In

Table 7 Interreviewer agreement scores

Explanations	Reviewer 1 vs. reviewer 2	Reviewer 1 vs. reviewer 3	Reviewer 2 vs. reviewer 3	All reviewers
All	0.87	0.49	0.33	0.57
Actual	0.82	0.24	0.01	0.39
Fabricated	0.93	0.70	0.63	0.75

Note: agreements between pairs of reviewers show the Cohen's kappa coefficient and agreement across all reviewers show the Fleiss' kappa statistic.

Table 8 Agreement rates for LIME-generated explanations, stratified by type of explanations and outcomes

Type of explanations and outcomes	Agreement rate (no. of agreements/no. of features)
Cases with actual explanations	
where patients had a dire outcome	0.81 (73/90)
where patients did not have a dire outcome	0.76 (68/90)
all patients	0.78 (141/180)
Cases with fabricated explanations	
where patients had a dire outcome	0.27 (8/30)
where patients did not have a dire outcome	0.77 (23/30)
all patients	0.52 (31/60)

Abbreviation: LIME, local interpretable model-agnostic explanations.

support of the third reviewer's judgment, commonly used scoring systems for assessment of severity of CAP, such as the pneumonia severity index[13] and CURB-65[29] do not include headache as a predictive feature. In the dataset, we used, almost all models included headache as a predictive feature; this may be because the Pittsburgh portion of the PORT data that we used in our experiments may have predictive features, such as headache, that are specific to the region. This indicates that predictive features in a model depend on the dataset that is used and explanations may uncover and inform physicians of features that are locally valid. More generally, this may suggest that predictive models should be derived from data that is from the location where the models will be deployed.

It is plausible that explanations of predictions are likely to be useful in clinical decision making,[10,11] and model-independent methods like LIME provide a method to generate explanations from any type of model.[15] However, it needs to be established that such explanations are valid, accurate, and easily grasped by physicians in the context of clinical predictive models. This study provides a first step toward that goal.

Limitations and Future Directions

This study has several limitations. Though LIME has the advantage that it can be used in conjunction with any predictive model, it has the limitation that internally it constructs a simple model. LIME constructs a local linear model from data in the neighborhood of the case of interest, and it seems reasonable to assume linearity in a small region even when the primary model is not linear. However, we and other investigators have noticed that the prediction of LIME's local model is not always concordant with the prediction of the primary predictive model.[30] Methods like LIME will need to be modified such that the predictions agree with those of the primary predictive model and work is ongoing in the research community to improve LIME.

This study used a single dataset that is relatively old (data collection occurred in the early 1990s), measures only one medical condition, and is limited to patient visits at a single geographical location. Additionally, the number of physician evaluators was relatively small. To explore the generalizability of using LIME with predictive models, newer datasets are needed in which different outcomes are measured and samples are collected from diverse geographical locations. Higher numbers of physician evaluators can also yield more reliable evaluations.

Conclusion

This study demonstrated that it is possible to generate patient-specific explanations to augment predictions of clinical outcomes by using available machine learning methods for both model development and generation of explanations. Moreover, explanations that were generated for predicting dire outcomes in CAP were assessed to be valid by physician evaluators. Such explanations can engender trust in the predictions and enable physicians to interpret the predictions in the clinical context.

Clinical Relevance Statement

This study demonstrated that there was good agreement among physicians on patient-specific explanations that are generated to augment predictions from machine learning models of clinical outcomes. Such explanations will enable physicians to better understand the predictions and interpret them in the clinical context, and might even influence the clinical decisions they make. Computerized clinical decision-supporting systems that deliver predictions can be enhanced to provide explanations to increase their utility.

Protection of Human and Animal Subjects

All research activities reported in this publication were reviewed and approved by the University of Pittsburgh's Institutional Review Board.

Funding

The research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM012095. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the University of Pittsburgh.

Conflict of Interest
None declared.

References

1 Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(01):198–208
2 Rothman B, Leonard JC, Vigoda MM. Future of electronic health records: implications for decision support. Mt Sinai J Med 2012; 79(06):757–768
3 Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer; 2009
4 Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2017;1:11
5 Liu S, Liu S, Cai W, Pujol S, Kikinis R, Feng D. Early diagnosis of Alzheimer's disease with deep learning. IEEE 11th International Symposium on Biomedical Imaging (ISBI) 2014:1015–1018
6 Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6(01):26094
7 Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017; 6:311–316
8 Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal lab tests. Machine Learning for Healthcare Conference. 2016:73–100
9 Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. Machine Learning for Healthcare Conference. 2016:301–318
10 Caruana R, Kangarloo H, David J, Dionisio N, Sinha U, Johnson D. Case-based explanation of non-case-based learning methods. Proc AMIA Symp 1999:212–215
11 Reggia JA, Perricone BT. Answer justification in medical decision support systems based on Bayesian classification. Comput Biol Med 1985;15(04):161–167
12 Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. ACM Digital Library 2015:1721–1730
13 Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med 1997;336(04):243–250
14 Lipton ZC. The mythos of model interpretability. Available at: https://arxiv.org/pdf/1606.03490.pdf. Accessed August 30 2019.
15 Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: explaining the predictions of any classifier. ACM Digital Library 2016: 1135–1144
16 Kim B. Interactive and interpretable machine learning models for human machine collaboration. PhD dissertation. Massachusetts Institute of Technology, 2015
17 Turner R. A model explanation system. IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2016:1–6
18 Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 2016;4(01):2
19 Štrumbelj E, Bosnić Z, Kononenko I, Zakotnik B, Kuhar CG. Explanation and reliability of prediction models: the case of breast cancer recurrence. Knowl Inf Syst 2010;24(02):305–324
20 Kapoor WN. Assessment of the Variantion and Outcomes of Pneumonia: Pneumonia Patient Outcomes Research Team (PORT) Final Report. Washington DC: Agency for Health Policy and Research (AHCPR); 1996
21 Cooper GF, Abraham V, Aliferis CF, et al. Predicting dire outcomes of patients with community acquired pneumonia. J Biomed Inform 2005;38(05):347–366
22 Caruana R. Iterated k-nearest neighbor method and article of manufacture for filling in missing values. United States Patent 6,047,287. May 5, 2000
23 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(Oct):2825–2830
24 Van Rijsbergen CJ. Information Retrieval. 2nd ed. Newton, MA, USA: Butterworth-Heinemann; 1979
25 Cohen J. A coefficient of agreeement for nominal scales. Educ Psychol Meas 1960;20:37–46
26 Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(05):378–382
27 Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics 1990;50(302):151–175
28 McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(03):276–282
29 Lim WS, van der Eerden MM, Laing R, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 2003;58 (05):377–382
30 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 2017:4765–4774
31 Krause J, Perer A, Ng K. Interacting with predictions: visual inspection of black-box machine learning models. ACM Conference on Human Factors in Computing Systems. 2016:5686–5697
32 Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Mueller K-R. How to explain individual classification decisions. J Mach Learn Res 2009;11:1803–1831
33 Sikonja MR, Kononenko I. Explaining classifications for individual instances. IEEE Trans Knowl Data Eng 2008;20:589–600
34 Štrumbelj E, Kononenko I. Towards a model independent method for explaining classification for individual instances. International Conference on Data Warehousing and Knowledge Discovery. 2008:273282
35 Lemaire V, Féraud R, Voisine N. Contact personalization using a score understanding method. Proceedings of the International Joint Conference on Neural Networks. 2008:649–654
36 Poulin B, Eisner R, Szafron D, et al. Visual explanation of evidence in additive classifiers. Proc Conference on Innovative Applications of Artificial Intelligence (IAAI06). 2006:1822–1829
37 Szafron D, Greiner R, Lu P, Wishart D, MacDonell C, Anvik J, et al. Explaining naïve Bayes classifications. Technical Report. Department of Computing Science, University of Alberta. 2003
38 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(03): 837–845
39 Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(01):77
40 Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927;22(158):209–212