Interpreting Black-box Models of Prognostic Risk by Stratified Linear Models

William R. Zame[1], Jinsung Yoon[1], Kartik Ahuja[1], Folkert W. Asselbergs[2,3,4] Mihaela van der Schaar[1,5,6*]

[1] University of California, Los Angeles, California, United States of America

[2] Farr Institute of Health Informatics Research and Institute of Health Informatics, University College London, London, United Kingdom

[3] Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom;

[4] Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, University of Utrecht, the Netherlands;

[5] University of Oxford, Oxford, United Kingdom

[6] Alan Turing Institute, London, United Kingdom

*mihaela.vanderschaar@eng.ox.ac.uk

Abstract

Background

Machine Learning models have been shown to provide better predictions than clinical risk scores used in medical practice. However, machine learning models are difficult to interpret and hence have not been widely accepted. This paper proposes a new method that improves the interpretability of machine learning methods, identifies risk predictors that are not used in existing risk models, and provides insights into the differing importance of individual patient features at different levels of risk.

Methods and Findings

Our new methodology (SLIM, stratified linear models) uses a given machine learning model to divide the population into risk strata and fits linear models (linear regressions and/or Cox proportional hazards models) within these risk strata. These linear models provide interpretations of the underlying machine learning model; in particular, the coefficients reflect the absolute and relative importance of different features for determination of risk in different strata .

Using the MAGGIC heart failure dataset and the UK Cystic Fibrosis Registry, we demonstrate that our method provides useful interpretations of a variety of black-box machine learning models. In particular, we identify patient features that are more/less predictive within each risk stratum; these are significantly different across risk strata. In some cases, this confirms existing clinical knowledge; in other cases it provides novel insights in risk prediction. We also demonstrate that the interpretations produced by our method are significantly more accurate than those produced by previous methods.

Conclusion

We show that existing machine learning models can significantly outperform simple regressions and existing clinical risk scores and still be interpreted in ways that are understandable to clinicians, confirm existing clinical knowledge and add new clinical variables to the prediction model.

Introduction

A substantial literature [1, 2] argues that Machine Learning (ML) models are capable of providing better predictions - e.g., better prognostic risk scores - than clinical models that are commonly used in many medical domains. Despite this, ML models have not gained wide acceptance. Perhaps the most important reason for this is that common clinical models are readily interpretable but that ML models are "black-boxes" that are *not* readily interpretable and hence are not trusted by clinicians, patients or medical researchers. This lack of acceptance of ML models has prompted a recent body of work on methods for interpreting Machine Learning models - but that work does not seem to have convinced medical researchers or clinicians [3–11]. The main objective of this study is to propose a new method of interpretation that is simpler and more readily understandable than existing methods, and to demonstrate - on the basis of large medical datasets - that this method is superior to previous methods of interpretation and can produce clinically useful insights and discoveries.

Our method begins with a space of patient features/covariates, a dataset of observations and a black-box ML model that predicts the probability that a patient with a particular set of features will experience a given event (e.g. death or onset of a disease within a given time horizon). We use the predictions of the given ML model to partition the space of features into a collection of *risk strata* (disjoint regions). The particular risk strata can be specified by the user (clinician, policy-maker, health economist, etc.) according to criteria deemed appropriate for the intended purpose. We then use the given dataset to fit a linear model (either a linear regression or a Cox proportional hazards model) to the ML on *each stratum* and patch these linear models together to create a *stratified linear model (SLIM).*

Many black-box models have the virtue that they are capable of incorporating hundreds of features - but human beings have a hard time *understanding* the impact of so many features (indeed most clinical models rely on a relatively small number of features even when many more are actually available). Moreover, some features are actionable and some are not: a clinician can treat the patient with statins or help the patient to lose weight, but age cannot be changed. It is therefore important to identify those features that are most/least important - and our approach does this. We show

that these features are often *different* across risk strata. In some cases, these findings reflect current clinical knowledge; in other cases our findings appear to represent new clinical discoveries.

To validate our approach we use two large medical datasets: the Meta-analysis Global Group in Chronic Heart Failure (MAGGIC) dataset, which consists of patients who have already experienced heart failure, and the UK Cystic Fibrosis Registry, which provides data for patients who have been diagnosed with Cystic Fibrosis (CF); in the first case the event of interest is death within 1 year, in the second case the event of interest is death or lung transplantation. In the main text, we focus on the MAGGIC dataset, relegating the description of and results for the CF dataset to the Supplementary Materials. We apply four different ML models (Random Forest, XGBoost, Gradient Boosting Machine (GBM) and a multi-layer Neural Network (NN)) that have been shown to predict well on these and other medical datasets [1, 2]. We demonstrate the predictive performance of these ML models on each of the datasets; the best ML models predict better than either conventional predictive models (linear regressions and Cox proportional hazards regression) or the best existing clinical model. For each of the ML models, we construct the corresponding SLIM and show how to use this SLIM to interpret the underlying ML model. We show that our method provides better fits - both in terms of error and in terms of consistent ranking - than previous methods of interpretation (regression trees, associative classifiers and a natural adaptation of LIME [11] to our setting).

Materials

Study Design

We conducted our study using two datasets. In the main text we focus on the MAGGIC dataset, which we describe below; in the Supplementary Materials we focus on the UK Cystic Fibrosis Registry, which is described and analyzed in the Supplementary Materials.

The MAGGIC dataset [12] is a collection of 30 different datasets from 30 different medical studies containing patients who experienced heart failure. Because the event of interest to us is death within 1 year, we excluded patients who were censored - disappeared from follow-up - before one year. The dataset provides 30,389 uncensored patients, of whom 5,723 (18.8%) patients died within the 1-year period. Among the 31 features/covariates provided in the MAGGIC dataset, we exclude the "Caucasian" feature because information on this feature is missing for too many patients (more than 10%). Because ACE-Inhibitors and ARB have the same effects, we do not distinguish between them but instead combine them into a single feature. Thus we are left with 29 features of which 20 are binary and 9 are continuous. Categorical binary features (e.g., Male/Female) are represented as 0, 1; other features are represented as real numbers. When information is missing, we use standard imputation methods to fill in the missing information. More specifically, we conduct 10 multiple imputations using Multiple Imputation by Chained Equations (MICE) as in [13].

Methods

We are given a space $\mathcal{X}$ of patient covariates/features and a set of possible labels/outcomes. In general, some of the features will be continuous and others will be categorical; without (much) loss of generality we assume all the categorical features are binary (e.g., gender) and represented by $0, 1$ and that the continuous features (e.g. creatinine) are represented by real numbers, so $\mathcal{X} \subset \mathbb{R}^D$. For simplicity, we assume that the only possible outcomes are $0, 1$: the label/outcome is 1 if the patient experienced the adverse event under consideration and 0 otherwise. We are also given a dataset.

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \quad (1)$$

where $\mathbf{x}_i \in \mathcal{X}$ describes the features of patient i and $y_i \in \{0, 1\}$ is the outcome experienced by patient i. Finally, we are given a black-box ML model $f : \mathcal{X} \to [0, 1]$; we interpret $f(\mathbf{x})$ as the predicted probability that a patient with features $\mathbf{x}$ experienced the adverse event. Tacitly, we assume that the dataset $\mathcal{D}$ was drawn i.i.d. from the true distribution on $\mathcal{X} \times \{0, 1\}$ and that the black-box ML model f was produced by training some algorithm on this dataset, but we make no explicit use of these assumptions.

The user (clinician, policy-maker, health economist, etc.) who wishes to use the interpretive model specifies a partition of $[0, 1] = \mathcal{Y}$ into subintervals:

$$\mathcal{Y} = \mathcal{Y}_1 \cup \ldots \cup \mathcal{Y}_K$$

where, by definition, $\mathcal{Y}_j \cap \mathcal{Y}_k = \emptyset$ if $i \neq j$. For convenience we agree to number subintervals from left to right - so that if $j < k$ then every point in $\mathcal{Y}_j$ precedes every point in $\mathcal{Y}_k$ - and that each subinterval is open on the left and closed on the right. For each subinterval $\mathcal{Y}_k$, write

$$\begin{aligned} \mathcal{D}_k &= \{(\mathbf{x}_i, y_i) \in \mathcal{D} : f(\mathbf{x}_i) \in \mathcal{Y}_k\} \\ \mathcal{X}_k &= \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \in \mathcal{Y}_k\} \end{aligned}$$

Thus $\mathcal{Y}_k$ is an interval of risk, $\mathcal{D}_k$ consists of those data points for which *predicted risk*

according to f lies in the interval $\mathcal{Y}_k$ and $\mathcal{X}_k$ consists of those covariates for which predicted risk according to f lies in the interval $\mathcal{Y}_k$; we call the regions $\mathcal{D}_k, \mathcal{X}_k$ the *risk strata*. It will be clear from context whether we are referring to the dataset or the feature space. It is important to keep in mind that the risk strata are defined by - and so depend on - the predictive model f.

For each k, we define $q_k^* : \mathcal{X}_k \to \mathcal{Y}$ to be the linear model (i.e. either a linear regression or a Cox proportional hazards regression) that best fits f on $\mathcal{D}_k$, in the sense of minimizing the mean squared error between the linear model and f on the dataset $\mathcal{D}_k$. If we write $\mathcal{L}$ for the set of all linear models, then q_k^* is formally defined by

$$q_k^* = \arg\min_{q \in \mathcal{L}} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_k} \Big[(f(\mathbf{x}) - q(\mathbf{x}))^2 \Big] \tag{2}$$

where the expectation is taken over the empirical distribution of $\mathcal{D}_k$. Note that we are minimizing the expectation of $(f(\mathbf{x}) - q(\mathbf{x}))^2$ and not of $(y - q(\mathbf{x}))^2$ because we are fitting to the given model f and not to the data. We define the *stratified linear model (SLIM)* $q^* : \mathcal{X} \to \mathcal{Y}$ by

$$q^*(\mathbf{x}) = \sum_{k=1}^{K} \Big[\mathbb{I}(\mathbf{x} \in \mathcal{X}_k) \times q_k^*(\mathbf{x}) \Big] \tag{3}$$

That is: we set $q^*(\mathbf{x}) = q_k^*(\mathbf{x})$ if $\mathbf{x} \in \mathcal{X}_k$. Although we have suppressed it to avoid clutter, it is important to keep in mind that q^* depends on the underlying model f and on the dataset $\mathcal{D}$. In fact, the risk strata $\mathcal{D}_k$ themselves depend on the underlying model and the dataset $\mathcal{D}$, as can be seen from Figure 2 in the Supplementary Materials.

According to the needs of the user, many criteria for defining these subintervals are possible. For example:

- The user prescribes sub-intervals so that the risk strata $\mathcal{D}_k$ partition the patients in the dataset into K equal-sized populations according to increasing risk (as predicted by f). In the Results section, we take $K = 5$ so that we are partitioning the dataset into population quintiles according to increasing risk. Such a partition would seem especially natural in situations where resources are constrained so that only a fraction of patients can be treated.

- The user prescribes sub-intervals corresponding to specific levels of risk. For

instance, the user might prescribe $\mathcal{Y}_1 = [0, 0.1), \mathcal{Y}_2 = [0.1, 0.2), \mathcal{Y}_3 = [0.2, 1.0]$ so that the Low Risk stratum $\mathcal{X}_1$ consists of patients whose predicted risk (according to f) is less than 10%, the Middle Risk stratum $\mathcal{X}_2$ consists of patients whose predicted risk is in the range 10% - 20%, and the High Risk group$\mathcal{X}_3$ consists of patients whose predicted risk is above 20%. Such a partition would seem especially natural in situations where it is understood that intervention might be appropriate only for patients with a particular level of risk.

Note that in each case, although we partition the entire space of features into risk strata, the user may be particularly interested in only one or two strata; e.g., perhaps only the patients at highest risk.

Metrics

Performance Metrics for ML Models

The clinical utility of a prognostic model should be evaluated in terms of the model's ability to distinguish patients who are truly at risk (perhaps in anticipation of intervention or treatment) from patients who are truly not at risk. Following standard usage, we consider three metrics of performance: the concordance index (C-index), the area under the receiver operating curve (AUROC), and the area under the precision recall curve (AUPRC). For the (standard) definitions, see the Supplementary Materials. We note that, if the true distribution is known and ties are irrelevant, the C-index and the AUROC are equal. In our actual datasets the C-index and the AUROC differ by less than 0.00001; since we report only 4 decimal places, we identify the C-index and the AUROC. In evaluating the performance of all the predictive models (the clinical models, the simple regressions and the ML models), we report both the C-index/AUROC and the AUPRC.

Performance Metrics for Interpretations

We report two metrics that quantify the fidelity of an interpretation Q to the original black-box model f. The first metric is the probability that they rank a randomly chosen pair of patients in the same way: $P(Q(\mathbf{x}) > Q(\mathbf{x'})|f(\mathbf{x}) > f(\mathbf{x'}))$. Empirically: among all pairs $(\mathbf{x}, \mathbf{x'}) \in \mathcal{D} \times \mathcal{D}$ of patient features for which $f(\mathbf{x}) > f(\mathbf{x'})$ we compute the

fraction for which $Q(\mathbf{x}) > Q(\mathbf{x'})$. This metric might be thought of as analogous to computing the C-index of Q when we treat the ML model f *as if* it were the true data (rather than a prediction) - keeping in mind that $f(x)$ is a probability not an outcome - so we call it the $\widehat{C}$-index.

The second metric is a direct measure of the error between an interpretive model Q and the ML model f. For an individual observation $(\mathbf{x}, y) \in \mathcal{D}$ this error is $f(\mathbf{x}) - Q(\mathbf{x})$. Note that, because we are trying to quantify the fidelity of the interpretation to the model, we are measuring the error between the interpretation and the model - not the error between the interpretation and the true outcome. As is usual, we regard large errors as more important that small errors so we use the squared error $[f(\mathbf{x}) - Q(\mathbf{x})]^2$; in evaluating over the entire data set or a single stratum we average, to produce the mean squared error $\mathbb{E}[f(\mathbf{x}) - Q(\mathbf{x})]^2$. As is usual, we report the square root of the mean squared error - i.e. the *root mean squared error* (RMSE) - rather than the mean squared error itself. Of course the mean squared error is just the square of the RMSE. Because the linear models we use (linear regression and Cox proportional hazards regression) are unbiased estimators of the underlying ML model, our interpretive model q^* (the constructed SLIM) has the property that the mean squared error is the variance of $f - q^*$ and the RMSE is the standard deviation of $f - q^*$. For each ML model and the constructed SLIM, we report the RMSE of $f - q^*$ both on the entire dataset and on each of the risk strata.

Results

We focus here on prognostic risk prediction for the MAGGIC dataset described in the Materials Section; the results for the CF dataset are described in the Supplementary Materials. We first provide the performance of the four predictive ML models (Random Forest (RF), Gradient Boosting Machine (GBM), XGBoost, and a multi-layer Neural Network (NN)), a simple linear regression, a Cox (proportional hazards) regression and the best-performing clinical model; the results are shown in Table 1 in terms of both the C-index/AUROC and the AUPRC. As can be seen in Table 1, the best ML models provide better performance than the best clinical models - but, because they are black-boxes, the ML models are harder to understand. As a sanity check, we computed calibration plots for the ML models to verify that they are all reasonably well calibrated; see the Supplementary Materials.

Model	C-index/AUROC	AUPRC
RF	0.7048 ± 0.0063	0.3625 ± 0.0113
GBM	0.7101 ± 0.0031	0.3729 ± 0.0116
XGBoost	0.7106 ± 0.0041	0.3706 ± 0.0111
NN	0.7034 ± 0.0021	0.3629 ± 0.0071
Linear Regression	0.6601 ± 0.0120	0.3077 ± 0.0114
Cox Regression	0.6955 ± 0.0037	0.3538 ± 0.0113
MAGGIC Score	0.6933 ± 0.0071	0.3423 ± 0.0121

Table 1. Predictive performance of machine learning models and clinical models

For purposes of illustration, we choose to partition according to population quintiles. Following the procedure described in the Methods Section, we then fit a SLIM to each of the ML models. Table 2 shows how well the SLIM fits each ML model, using both the $\widehat{C}$-index and RMSE.

Model	RMSE	$\widehat{C}$-index
RF	0.0705	0.9198
GBM	0.0388	0.9406
XGBoost	0.0298	0.9378
NN	0.0310	0.9469

Table 2. RMSE and $\widehat{C}$-index of SLIM for ML models - MAGGIC dataset

Because the fit varies substantially across risk strata, we also show in Table 3 the RMSE fits in each population quintile.

Model	Risk Strata				
	0 - 20%	20% - 40%	40 - 60%	60 - 80%	80 -100%
RF	0.0084	0.0075	0.0097	0.0194	0.1553
GBM	0.0106	0.0134	0.0177	0.0270	0.0759
XGBoost	0.0168	0.0143	0.0158	0.0215	0.0552
NN	0.0095	0.0120	0.0146	0.0234	0.0571

Table 3. RMSE of SLIM for ML models for each risk group - MAGGIC Dataset

For the best-performing ML models (GBM and XGBoost) we show the coefficients of the linear model within each risk stratum in terms of heat maps: Figures 1 and 2.

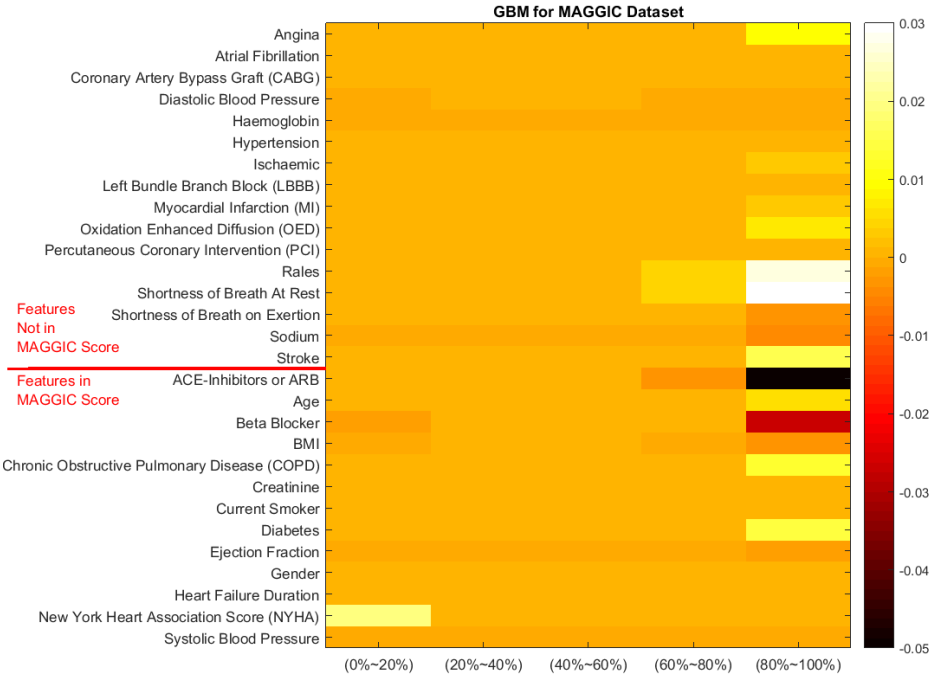

Fig 1. Coefficient of each feature for each population quintile in MAGGIC dataset - GBM

How to Read the Heat Maps

The heat maps show the importance of different features (as measured by the value of the coefficients of the linear model) within each risk stratum. Lighter colors represent features that have a higher positive predictive value; darker colors represent features that have a higher negative predictive value; colors in the middle have the least predictive value.

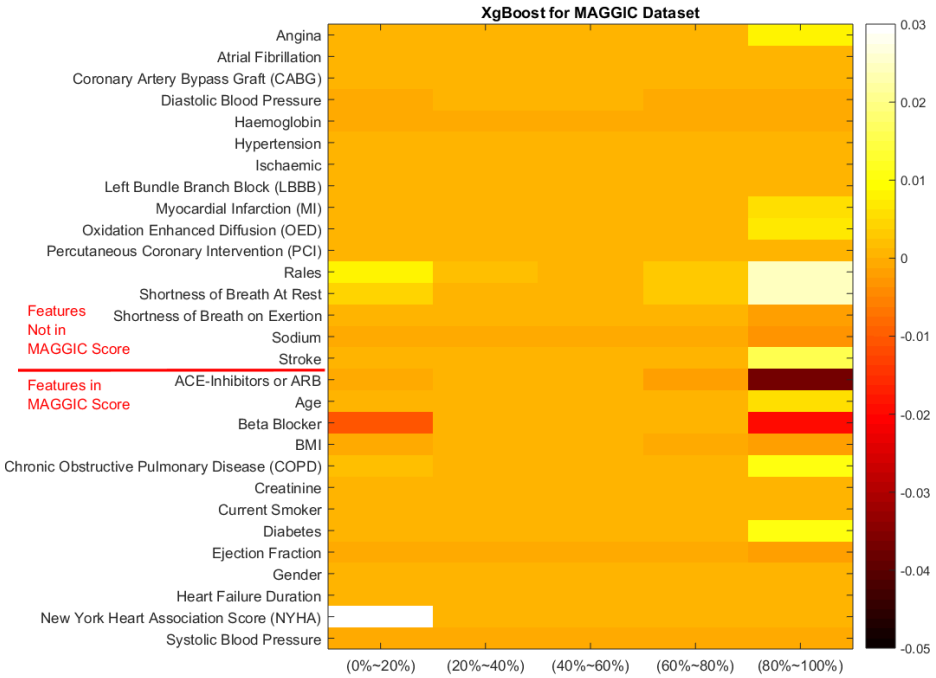

Fig 2. Coefficient of each feature for each population quintile in MAGGIC dataset - XGBoost

From these heat maps we can draw a number of inferences. (1) The heat maps for GBM and for XGBoost are similar; this means that GBM and XGBoost have similar views of the importance of most features within each population quintile. For instance, both GBM and XGBoost view age as a very important factor within the lowest and highest quintiles but as a much less important factor within the middle quantiles. (2) In some cases, the heat maps confirm prior medical knowledge. For example, the heat maps confirm the known predictive importance of the New York Heart Association (NYHA) score within the lowest risk quintile. (3) However, some care must be taken when interpreting the importance of a given feature within a particular risk quintile. For example, the NYHA score is not very predictive within the higher risk quintiles for the simple reason that there is very little variation of the NYHA score within these quintiles; e.g., as Table 4 shows, within the highest risk quintile (as predicted by either GBM or XGBoost), more than 99% of patients have NYHA score of 3 or 4 (treated as high). By contrast, within the lowest risk quintile (as predicted by either GBM or XGBoost), there is substantial variation of the NYHA score: more than 30% of patients have

NYHA score of 1 or 2 (treated as low), while the remaining patients have NYHA score of 3 or 4. (4) In some cases, the heat maps provide new medical information (discovery). For example, we see that for the middle quintiles (the patients whose risk places them within the 20th-80th percentile), most of the features are equally (but only weakly) predictive. In particular, the 13 features used in the MAGGIC Score are no more (or less) predictive than the other 16 features. (5) On the other hand, the heat maps highlight that a relatively small number of features are most predictive for the highest risk group (patients above the 80th percentile) even though these features are not highly predictive for other risk groups. For example, Rales and Shortness of Breath - features that are *not* included in the MAGGIC score - are in fact the most predictive features for patients in the highest risk quintile but are much less predictive in other risk quintiles.

Dataset	Model	0% - 20%	20% - 40%	40% - 60%	60% - 80%	80% - 100%
MAGGIC	RF	72.3%	89.6%	94.0%	97.6%	96.7%
	GBM	64.5%	91.6%	96.2%	98.6%	99.5%
	XGBoost	68.8%	90.3%	94.6%	97.6%	99.1%
	NN	76.0%	88.6%	92.7%	95.3%	97.9%

Table 4. NYHA Distribution for each risk group

Alternative Methods of Interpretation

Our method provides interpretations of the various ML models; to provide some context we compare the quality of our interpretations with other methods of interpreting ML models: Associative Classifiers, regression trees and an adaptation of LIME to our setting.

Associative Classifiers are described in detail in [14, 15]. To apply the method in our context, we first divide the dataset into quintiles according to population - exactly as we have done in producing risk strata. We then divide each population quintile into deciles (called classes) according to risk; e.g. if the lowest risk group consists of patients whose risk is in the range $[0.0, 0.1]$, the deciles are $[0.0, 0.01], (0.01, 0.02], \ldots, (0.09, 0.1]$. Each of these deciles is a class. We then use the data to learn the correlations between the classes and the given features (or conjunctions and disjunctions of features); [14] provides an algorithm for transforming these correlations into a predictive model, which is viewed as an interpretation of the given ML model.

Alternatively, we can use regression trees in place of linear models to produce interpretations of the underlying ML model within each risk stratum. We first divide the dataset into quintiles according to population. Within each quintile (each risk stratum), we grow a regression tree [16]. That is, we use a feature to split the risk stratum into two nodes for which the outcomes are most homogeneous, and repeat the process within each node. In order to retain a manageable interpretation we restrict the depth of the tree to be at most 3, so there are at most 8 leaves (terminal nodes). Within each leaf, we predict the probability of the adverse event as the ratio of the number of patients in that leaf who experienced the event to the total number of patients in that leaf. This yields a single predictive model within each risk stratum.

Interpretation by means of associative classifiers and by means of regression trees are similar. Both produce interpretations that are piecewise constant within each risk stratum, and neither yield information about the relative importance of features within each risk stratum.

LIME [11] (in the natural adaptation we use) is closer to our approach. The version of LIME described in [11] is intended to interpret classification models rather than regression models; since we follow the spirit of the method rather than the letter, we call the interpretation we use LIME-R. LIME-R again begins by dividing the dataset into quintiles by population. For each risk stratum and patient (in the dataset) in that risk stratum, LIME-R creates a linear regression that approximates the original ML model in a neighborhood of that patient. We aggregate these linear regressions across the entire risk stratum to provide a single linear regression whose predictions can be compared to those of the original ML model across the entire risk stratum. Moreover, as is the case with our method, the coefficients of this linear regression can be interpreted as representing the importance of various features.

LIME-R differs from our method in two important ways. The first is that LIME-R necessarily produces a linear regression within each stratum whereas our approach sometimes (in fact, often) produces a Cox proportional hazards regression - according to whichever is more accurate. It is evident that for some datasets (and especially for the highest risk stratum) the truth - and the predictive ML model - might most closely resemble a Cox regression so that forcing the interpretation to be a linear regression will create a poorer fit. The second is that LIME-R produces its linear regression by

averaging local linear regressions, rather than by any optimizing procedure; this also leads to poorer fit.

Comparing Interpretations

We compare our interpretative method with each of the others described above in several ways. In Table 5 we report the $\hat{C}$ index and the RMSE (computed over the entire dataset $\mathcal{D}$) for our interpretation and for the other methods. A higher $\widehat{C}$ index and a lower RMSE represent better fits, so we see that, for each of the ML models, our interpretive method performs better (produces interpretations that fit the underlying model better) - and, for all of the ML models except Random Forest, much better - than any of the other interpretive methods.

Note that all four interpretive methods do much more poorly on (the interpretation of) Random Forest than on the other ML models, so a clinician who chooses among ML models on the basis of *both* predictive accuracy *and* interpretability would therefore downgrade Random Forest on this basis.

Metrics	Model	SLIM	Regression Tree	Associative Classifier	LIME-R
RMSE	RF	0.0705	0.0742	0.0745	0.0736
	GBM	0.0388	0.0533	0.0551	0.0512
	XGBoost	0.0298	0.0409	0.0423	0.0391
	NN	0.0310	0.0496	0.0515	0.0480
$\widehat{C}$-index	RF	0.9198	0.8663	0.8395	0.8868
	GBM	0.9406	0.8842	0.8478	0.9137
	XGBoost	0.9378	0.8815	0.8445	0.9103
	NN	0.9469	0.8818	0.8504	0.9216

Table 5. RMSE and $\widehat{C}$-index of SLIM for ML models in comparison to the benchmarks

To explore further, we show in Table 6 the RMSE error comparison of our method with LIME-R (which is the best of the other interpretive methods) within each risk stratum. As can be seen, our method performs better than LIME-R across all ML models and risk strata, and, again with the exception of Random Forest (where both our method and LIME-R do less well), significantly better in all strata and much better in most strata.

We do not compare the $\widehat{C}$-index within each risk stratum because the variance of risk within each stratum is small - especially in the lower strata. This means that two

randomly chosen patients are very likely to have similar risk scores, so that preserving the risk ranking within a stratum is a much harder task than preserving the ranking across the entire population, and it is therefore harder to interpret $\widehat{C}$-index within each risk stratum. Moreover, because the variance of risk within the lower risk strata is small, mis-ranking within a risk stratum is likely to be much less important than mis-ranking across the entire population.

Interpreter	ML Model	Risk Strata				
		0% - 20%	20% - 40%	40% - 60%	60% - 80%	80% -100%
SLIM	RF	0.0084	0.0075	0.0097	0.0194	0.1553
	GBM	0.0106	0.0134	0.0177	0.0270	0.0759
	XGBoost	0.0168	0.0143	0.0158	0.0215	0.0552
	NN	0.0095	0.0120	0.0146	0.0234	0.0571
LIME-R	RF	0.0088	0.0076	0.0095	0.0212	0.1556
	GBM	0.0143	0.0148	0.0189	0.0292	0.0991
	XGBoost	0.0218	0.0154	0.0165	0.0227	0.0706
	NN	0.0158	0.0146	0.0180	0.0286	0.1051

Table 6. RMSE of SLIM and LIME-R for ML models for each risk stratum - MAGGIC Dataset

We stress that, while a good fit to the underlying ML model would seem a necessary characteristic of any interpretation, it is not the only desirable characteristic; we also want the interpretation to provide information about the importance of features. Both SLIM and LIME-R provide reasonably fine information, but associative classifiers and regression trees - by their nature - provide only coarse information.

Discussion

In this study, we develop a methodology for interpreting the predictions of ML models of prognostic risk prediction. Our method and associated findings are important because the lack of interpretability of ML models represents an important obstacle to their acceptance and use by the medical community - even in settings where these ML models have demonstrated predictive performance that is superior to that of clinical models. Our method and associated findings are also important because they provide confirmation of existing clinical knowledge of the relationships between covariates and risk and also discover new information about these relationships. Our method is more flexible and easier to understand than previous methods.

In the original paper describing the MAGGIC risk score the authors used 13 independent predictors (features). Our interpretive models confirmed the predictive value of these 13 features but also demonstrated that an additional 16 features are equally or more relevant for best predictions. All of these 29 features are factors that have been known to relate to the progression of heart failure, although their relative importance has perhaps not been understood. One advantage of the machine learning approach is that in can easily adapt to the needs of the clinician by highlighting the features that are actionable such as co-morbidities and medication use. As shown in the stratified analyses based on risk, differences in importance of particular features exist across the different strata. In the era of precision medicine, models that provide individualized risk prediction are needed, not just one-size-fits-all models that do not seem to apply for common diseases.

Another advantage of machine learning models is that they are capable of handling many features - as opposed to existing rigid clinical models that make use of a relatively small set of features. This is becoming even more important as electronic health records - which provide enormous amounts of information - become more and more widely available, because electronic health records provide administrative data, patient reported data, clinical examination data, measurement data, imaging data, laboratory (biomarker) data, and medication data far beyond what is commonly used by existing clinical models.

Key findings

We emphasize several key findings of our study:

- Our method is competitive with, and usually outperforms, existing methods of interpreting machine learning methods for prognostic risk prediction.
- Our method demonstrates that the *relative importance* of patient features is different for different risk strata in the patient population. This is especially important in guiding treatment choices/decisions.
- Our method confirms medical knowledge of the importance of some features but also identifies other features whose importance was not known (or at least not incorporated into the leading clinical risk scoring methods).

Limitations

Our work has several important limitations. The first is that the interpretation can only be as good as the ML model it is interpreting. The second is that there is no guarantee that our method will produce a good interpretation of *every* ML model (and indeed it does not). But this limitation is, in a way, also a virtue, because it provides a way of choosing *among* ML models on the basis of interpretability and not just on the basis of performance. The third is that we have not performed a validation study using an external dataset as is done in classical epidemiological studies. The fourth is that, although our work demonstrates the added value of machine learning methods for risk prediction, the value of machine learning methods depends, just as classical regression analyses and clinical models do, on the availability of data. In routine clinical practice it is often the case that information is missing: measurements were not taken or not recorded. In such settings it is necessary either to impute the missing information before applying the risk algorithm, or to include missingness in the algorithm, which requires making assumptions about why certain information is missing. For instance, certain measurements may not have been taken because the clinician believed, on the basis of other measurements, that they would not be relevant.

Conclusion

We offer a new method for interpreting the risk predictions of black-box machine learning models. Our method is explicitly designed to address the heterogeneity of patient populations and the needs of users (clinicians, policy-makers, health economists, etc.) by identifying various sub-populations according to risk. Our method captures the different effects of covariates and interactions between covariates for these various sub-populations. Our method out-performs previous methods of interpretation while being more flexible and more easily understandable to clinicians.

Acknowledgments

Jinsung Yoon is supported by the U.S. National Science Foundation and Office of Naval Research.

Folkert Asselbergs is supported by UCL Hospitals NIHR Biomedical Research Centre and EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking BigData Heart grant number 116074.

Mihaela van der Schaar is supported by the U.S. National Science Foundation and Office of Naval Research.

Author contributions

William Zame, Jinsung Yoon, Kartik Ahuja and Mihaela van der Schaar contributed equally. Folkert Asselbergs contributed medical expertise.

Code and Data Sharing

MAGGIC and UKCFR data are accessible through a request process. Codes used for the analysis are publicly available in `https://github.com/jsyoon0823/SLIM`.

References

1. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems. 2012;36(4):2431–2448.

2. Yoon J, Zame WR, Banerjee A, Cadeiras M, Alaa AM, van der Schaar M. Personalized survival predictions via Trees of Predictors: An application to cardiac transplantation. PloS one. 2018;13(3):e0194985.

3. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one. 2015;10(7):e0130140.

4. Bastani O, Kim C, Bastani H. Interpreting blackbox models via model extraction. arXiv preprint arXiv:170508504. 2017;.

5. Cook RD, Weisberg S. Residuals and influence in regression. New York: Chapman and Hall; 1982.

6. Koh PW, Liang P. Understanding black-box predictions via influence functions. arXiv preprint arXiv:170304730. 2017;.

7. Lakkaraju H, Kamar E, Caruana R, Leskovec J. Interpretable & Explorable Approximations of Black Box Models. arXiv preprint arXiv:170701154. 2017;.

8. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems; 2017. p. 4768–4777.

9. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition. 2017;65:211–222.

10. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? arXiv preprint arXiv:161107450. 2016;.

11. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 1135–1144.

12. Pocock SJ, Ariti CA, McMurray JJ, Maggioni A, Køber L, Squire IB, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. European heart journal. 2012;34(19):1404–1413.

13. Buuren Sv, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. Journal of statistical software. 2010; p. 1–68.

14. Johnson I. arulesCBA: Classification for Factor and Transactional Data Sets Using Association Rules;.

15. Antonie ML, Zaïane OR. An associative classifier based on positive and negative rules. In: Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM; 2004. p. 64–69.

16. Breiman L. Classification and regression trees. Routledge; 2017.

Supporting information

S1 Paragraph. Performance Metrics

S2 Paragraph. Materials: UK Cystic Fibrosis Registry

S1 Fig. Calibration plots for 4 different machine learning models on the MAGGIC dataset.

S2 Fig. The relationship between risk and cumulative percentile - MAGGIC dataset

S3 Fig. Calibration plots for 4 different machine learning models on the UKCFR dataset

S4 Fig. The relationship between risk and cumulative percentile - UKCFR dataset

S5 Fig. Coefficient of each feature for each risk stratum in UKCFR dataset - GBM

S6 Fig. Coefficient of each feature for each risk stratum in UKCFR dataset - XGBoost

S1 Table. Predictive performance of machine learning models and clinical models - UKCFR dataset

S2 Table. RMSE and C-index of SLIM for ML models - UKCFR dataset

S3 Table. RMSE and C-index of SLIM for ML models in comparison to the benchmarks - UKCFR dataset

S4 Table. RMSE of SLIM and LIME-R for ML models for each risk group - UKCFR dataset