Robust Local Explanations for Healthcare Predictive Analytics: An Application to Fragility Fracture Risk Modeling

Buomsoo (Raymond) Kim
University of Arizona, buomsookim@email.arizona.edu

Karthik Srinivasan
University of Arizona, karthiks@email.arizona.edu

Sudha Ram
University of Arizona, sram@email.arizona.edu

Robust Local Explanations for Healthcare Predictive Analytics: An Application to Fragility Fracture Risk Modeling

Buomsoo (Raymond) Kim
Eller College of Management,
University of Arizona
McClelland Hall, 1130 E Helen St,
Tucson, AZ 85721, US
buomsookim@email.arizona.edu

Karthik Srinivasan
School of Business,
University of Kansas
1654 Naismith Dr, Lawrence,
KS 66045, US
karthiks@ku.edu

Sudha Ram
Eller College of Management, University of Arizona
McClelland Hall, 1130 E Helen St, Tucson, AZ 85721, US
sram@email.arizona.edu

Abstract

With recent advancements in data analytics, healthcare predictive analytics (HPA) is garnering growing interest among practitioners and researchers. However, it is risky to blindly accept the results and users will not accept the HPA model if transparency is not guaranteed. To address this challenge, we propose the RObust Local EXplanations (ROLEX) method, which provides robust, instance-level explanations for any HPA model. The applicability of the ROLEX method is demonstrated using the fragility fracture prediction problem. Analysis with a large real-world dataset demonstrates that our method outperforms state-of-the-art methods in terms of local fidelity. The ROLEX method is applicable to various types of HPA problems beyond the fragility fracture problem. It is applicable to any type of supervised learning model and provides fine-grained explanations that can improve understanding of the phenomenon of interest. Finally, we discuss theoretical implications of our study in light of healthcare IS, big data, and design science.

Keywords: healthcare predictive analytics, explainable artificial intelligence, fragility fracture

Introduction

With recent developments in big data technology and the prevalence of large-scale datasets from diverse sources, the healthcare predictive analytics (HPA) field is witnessing a dramatic surge of interest. The primary objective of HPA is to predict health-related events based on patterns in data (Lin et al. 2017). Previous studies in HPA proposed novel methods to improve the prediction of health outcomes such as readmission of patients (Bardhan et al. 2015) and chronic conditions (Lin et al. 2017). Nevertheless, in healthcare, it is not only important to provide accurate predictions, but also critical to provide reliable explanations and rationale behind the model-based predictions. Such explanations can play a crucial role in not only supporting clinical decision-making but also facilitating user engagement and patient safety (Ahmad et al. 2018). If users and decision-makers do not have faith in the HPA model, it is highly likely that they will reject its use (Ribeiro et al. 2016). Furthermore, it is extremely risky to blindly accept and apply the results derived from black-box models, which might lead to undesirable consequences or life-threatening outcomes in domains with high stakes such as healthcare (Ahmad et al. 2018).

To accommodate the need for transparency in HPA, we propose a novel method called **RO**bust **L**ocal **EX**planations (ROLEX) to generate local explanations for complex prediction models. Our ROLEX method

differs from existing interpretable HPA methods (e.g., Che et al. 2016; Choi et al. 2016) since our focus is on generating robust model-agnostic local explanations – i.e., generalizable and fine-grained explanations. Model-agnosticity refers to the capability of flexibly explaining any supervised prediction model (Ribeiro et al. 2016). Existing methods in HPA attempt to explain specific prediction models such as recurrent neural networks or tree-based ensembles. Such approaches are not aligned with recent trends in HPA, in which a wide variety of datasets are used along with diverse types of prediction models. Model-agnostic explanation methods enhance the applicability of HPA to a wide range of problems. In HPA, it is common to encounter complex, non-linear relationships between variables. In such cases, it is often insufficient to explain the underlying phenomenon by providing only global, dataset-level explanations such as feature importance scores or coefficients. Local explanations refer to instance-level explanations, i.e., explanations focusing on a small part of the input space that can be effectively understood by humans (Ribeiro et al. 2016). Granular, instance-level explanations can help to facilitate personalized care. Finally, our ROLEX method is more robust than existing methods since it minimizes the influence of outliers and standardizes feature ranges with a min-max scaling procedure. This is particularly useful because datasets in healthcare and medicine are rife with outliers, skewed class distribution, and non-standardized feature ranges.

The applicability of our proposed ROLEX method is demonstrated using a specific case of HPA, viz., the fragility fracture prediction problem. Fragility fractures are defined as "fractures that result from mechanical forces that would not ordinarily result in fracture." (National Clinical Guideline Centre 2012) According to the International Osteoporosis Foundation (2018), around 2.7 million fractures occurred in six developed European countries during 2017, resulting in estimated costs exceeding €37.5 billion. Further, such economic costs are expected to grow by 27% in the next decade. The resulting medical and economic burdens are driving the compelling need for early diagnosis and prevention of osteoporosis and fragility fractures. The need for early identification of fractures has been recognized by the World Health Organization (WHO), leading to the development of the FRAX® model (Kanis et al. 2009), which became the de-facto method to evaluate fracture risks in practice. However, limitations of the FRAX® model, such as the exclusion of some clinically important factors, have been constantly suggested. Such factors include measurements of physical activity, vitamin deficiency, and previous fractures (Silverman and Calderon 2010). Thus, recent studies applied predictive modeling with extended features to overcome such limitations. However, extended feature sets and sophisticated algorithms often result in less transparency and interpretability in general. We address the need to enhance transparency and interpretability of fragility fracture risk prediction models by providing robust explanations with the ROLEX method. We show that the ROLEX method outperforms state-of-the-art methods (Laugel et al. 2018; Ribeiro et al. 2016) in terms of *local faithfulness* in generating local, model-agnostic explanations. Here, *local faithfulness* is defined as the ability to replicate the original classifier in the vicinity of the instance of interest measured by an objective metric defined mathematically. Finally, we derive a few key insights on fragility fracture risks across a highly heterogeneous patient population based on local explanations developed using the ROLEX method. Our key findings have theoretical implications for various streams of information systems (IS) research, including healthcare IS, predictive analytics, and design science, along with practical implications for physicians and patients.

The remainder of this paper is organized as follows. In the following section, we discuss related work from three research streams – (1) healthcare predictive analytics, (2) explainable artificial intelligence and machine learning interpretability, and (3) fragility fracture prediction. Next, we present our proposed method, i.e., the ROLEX method, and demonstrate its application to the fragility fracture prediction problem with a real-world dataset, and also empirically compare it with existing methods. Finally, we discuss key findings and limitations and provide directions for future work.

Related Work

Our study lies at the intersection of three highly relevant research streams – HPA, fragility fracture prediction, and explainable artificial intelligence and machine learning interpretability. In this section, we first describe previous works in HPA and how our study differs from existing works. Then, we review previous works in fragility fracture predictive modeling and discuss recent developments in explanation techniques for complex prediction models that set the foundation of our study.

Healthcare Predictive Analytics

The primary goal of HPA is to accurately predict health-related outcomes such as medical complications, treatment responses, and patient mortality. With an increase in the availability of large-scale data in the medical and healthcare fields – e.g., electronic health records (EHR), cohort studies, clinical trials, and mobile healthcare – there has been a dramatic growth in HPA (Chen et al. 2012; Lin et al. 2017). Previous studies in HPA have developed novel methods to accurately predict diverse health outcomes. For example, Lin et al. (2017) proposed a Bayesian multitask learning framework for risk profiling in chronic care, while Meyer et. al., (2014) integrated machine learning and control theory for chronic care.

Nevertheless, our study differs from previous studies in HPA in that we shed light on *how* HPA can support decision making by building transparent systems that stakeholders can understand and trust. Healthcare, with the wide-spread use of big data, is rife with applications warranting transparency and rationalization of quantitative model-based predictions (Ahmad et al. 2018; Choi et al. 2016). As a result, transparency and interpretability of machine learning-based decision support systems are vital in healthcare. This is especially critical when multidisciplinary teams frequently collaborate, entailing clear and unequivocal communication between professionals with diverse backgrounds such as surgeons, hospitalists, and medical coders. (Fichman et al. 2011). Furthermore, explanations play a vital role in user engagement and patient safety (Ahmad et al. 2018) for a variety of reasons. First, users generally require trust to employ the models in the wild – "*if the users do not trust a model or a prediction, they will not use it*" (Ribeiro et al. 2016, italics in the original). Providing explanations that are understandable by users is likely to enhance trust and promote acceptance of the model (Ribeiro et al. 2016). Lack of trust is a significant issue in the medical field, in which stakeholders are conservative and resistant to innovation and adoption of new technologies (Fichman et al. 2011). Second, blindly accepting and applying the results from a black-box model may lead to grave consequences, possibly life-threatening, in domains where the stakes are high (Ahmad et al. 2018; Doshi-Velez and Kim 2017). Some might contend that testing the model with a hold-out or cross-validated dataset is sufficient to evaluate its applicability. However, inherent biases in data (Executive Office of the President 2016) and the complex nature of real-world problems often call for extended human interpretation and intervention (Doshi-Velez and Kim 2017).

A few recent studies have explored the interpretability of prediction models in healthcare. Choi et al. (2016) and Che et al. (2016) both proposed HPA models for EHR using interpretable deep learning. Choi et al. (2016) utilized attention networks, which was initially used to align words in neural translation (Bahdanau et al. 2015), while Che et al. (2016) effectively apply gradient boosting trees. However, their frameworks are focused on explaining only one type of models, i.e., deep learning-based models, and one type of data, i.e., EHR data, with limited generalizability. Therefore, there is a need for interpretable predictive modeling systems that are generalizable to different healthcare-related problems and context.

Fragility Fracture Prediction

Fragility fractures are becoming a serious medical and economic issue worldwide, especially in developed countries (Reginster and Burlet 2006). It is reported that a 50-year-old female in Japan has a 40% probability of having a vertebral compression fracture during her lifetime (Tsuda 2017). In 2017, around 2.7 million fragility fractures occurred in six European countries alone, i.e., France, Germany, Italy, Spain, Sweden, and the UK. Fracture-related costs were estimated at €37.5 billion in 2017 and are expected to increase up to 27% by 2030 (International Osteoporosis Foundation 2018). With growing concerns over medical and economic burden of fracture-related diseases, the importance of early diagnosis and intervention before the first fracture occurrence is critical (Reginster and Burlet 2006). Furthermore, once individuals with high fracture risks are identified in advance, effective measures can be taken to prevent them from developing a fragility fracture. It has been demonstrated that active medication for women with osteoporosis can significantly reduce their fracture risks (Black et al. 2000), while regular exercise and extra care in dietary intake can reduce the risk of fracture in general (NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy 2001).

Accordingly, multiple studies in the past have focused on predicting fractures. FRAX® is arguably the most established analytical method to assess bone fracture risk; it not only makes risk assessments but also provides clinical guidelines. To estimate fracture risks, Poisson regression models are trained to assess hip and other osteoporotic fractures, with and without bone mineral density (BMD) scores. BMD refers to the

amount of bone mineral in bone tissue (US National Library of Medicine 2008) and it is a widely recognized predictor of fractures (Kanis et al. 2001). The FRAX® model was developed and validated using large-scale cohort-based datasets from Europe, North America, Asia, and Australia and provides 10-year probabilities of fracture (Kanis et al. 2009). Although it is an established method, FRAX® suffers from several limitations, including the exclusion of clinically important factors such as physical activity levels and vitamin deficiency, and failure to account for patient-level information such as ethnicity and socioeconomic factors (Silverman and Calderon 2010). Recent attempts to improve fracture prediction in the HPA community have used more extensive features and sophisticated prediction algorithms with recent advancements in data analytics. Atkinson et al. (2012) have used bone imaging data along with gradient boosting machines to assess fracture risks, while Iliou et al. (2014) have considered diagnostic factors – i.e., age, sex, height, and weight – to detect osteoporosis using machine learning. More recently, Forgetta et al. (2018) demonstrated the viability and utility of genomics-based prediction with machine learning methods. However, to the best of our knowledge, there is no large-scale HPA study effectively utilizing a multi-faceted, longitudinal dataset to predict the risk of bone fracture. Medical literature has pointed out that along with demographic factors and bone mineral density related information, extrinsic factors can significantly influence fracture risks. Such extrinsic factors include, but are not limited to, lifestyle, physical activity, nutrition (NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy 2001) and previous medical history (Silverman and Calderon 2010). Therefore, for a comprehensive analysis, such risk factors should be considered in an HPA model for fragility fracture risk assessment. Our cohort dataset, which we describe in detail in the subsequent sections, includes critical factors which were not previously considered. However, the inclusion of diverse features and utilization of sophisticated learning algorithms generally imply less transparency and interpretability in the prediction model. Thus, more effort should be invested in making the model deployable "in the wild" by enhancing human interpretability. Therefore, in the next section, we review relevant articles in explaining complex prediction models and suggest where we can contribute to HPA in real-world problem-solving settings.

Machine Learning Interpretability and Explainable Artificial Intelligence

As machine learning and artificial intelligence systems are becoming more capable and ubiquitous, *explainable artificial intelligence* and *machine learning interpretability* are garnering significant attention among practitioners and researchers (Doshi-Velez and Kim 2017). The introduction of policies such as the General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union 2016) has amplified the need for ensuring human interpretability of prediction models. Though there are no clear-cut standards for categorizing explanation and interpretation methods, we rely on two criteria that are widely accepted in the field – i.e., the scope of explanations and model specificity (Molnar 2019; Ribeiro et al. 2016). One criterion to categorize explanation methods is based on the scope of explanations. Global explanations attempt to figure out patterns at a dataset level. This is usually achieved by examining the model and the dataset *per se*. Coefficient estimates and *p*-values of input variables for a logistic regression model would be a classic example of global explanations (Molnar 2019). In contrast, local explanations focus on a part of the dataset – i.e., certain data instances of interest. Interactions between different features and relationships between input and outcome variables are likely to be highly non-linear in many datasets. Nonetheless, the relationship between the outcome(s) and inputs may be explained using a different linear function for each instance, or through instance-level explanations (Molnar 2019; Ribeiro et al. 2016). In other words, generating local, linear explanations can help understand local relationships in subsets of the high-dimensional, non-linear feature space. In the context of designing information technology (IT) artifacts for problem-solving, our interpretation of local explanations is essentially scaling down the "reducible uncertainty" and "loss of intellectual control" by decomposing a complex problem into smaller and simpler sub-problems (Rai 2017, p. 5-6). In his seminal work, Simon (1996) demonstrated that most complex systems are "nearly decomposable" and sub-components perform sub-functions that can contribute to the function of the whole system. Thus, one way to design a complex artifact is to find "viable ways of decomposing it into semi-independent components corresponding to its many functional parts." In a similar vein, local explanations for intelligent systems can be regarded as reverse-designing a complex artifact – e.g., a black-box prediction model - to understand the artifact and phenomenon of interest more deeply. Another criterion utilized to categorize explanation methods is model specificity. Explanation methods can either be considered as model-specific or model-agnostic. Model-specific methods are applicable to only a certain type of models. In many cases, model-specific methods pertain to one of the inherent properties of the model. For instance, coefficient estimates and *p*-values would be model-specific

methods applicable to the family of linear models. In contrast, model-agnostic methods are flexible enough to be applied to different types of supervised learning models. In other words, the model is regarded as a black box, regardless of the model structure and learning algorithms (Ribeiro et al. 2016). For most model-agnostic methods, a black-box prediction model – e.g., deep neural networks or an ensemble model - is trained in advance. Then, an interpretable surrogate model – e.g., logistic regression or decision trees – is trained to explain the black-box model.

Combining the two criteria results in four types of explanation methods – i.e., (1) global, model-specific methods, (2) global, model-agnostic methods, (3) local, model-specific methods, and (4) local, model-agnostic methods. We focus on *local, model-agnostic* methods in this study to develop an explanation method that is suitable for solving many HPA problems. Variety, along with sheer volume and velocity, are three important characteristic feature of datasets in the big data era (Laney 2001). The healthcare domain is no exception with a number of data sources that are distinctly different from each other, e.g., EHR, mobile applications, and sensors (Chen et al. 2012). Furthermore, datasets from different sources often require distinct prediction models. For instance, image data from large-scale radiology studies require models optimized for image processing such as convolutional neural networks. Whereas for clinical text data other types of models such as hidden Markov models and recurrent neural networks may be more effective. Thus, to maximize the applicability to any HPA problem in a timely manner, proposing a model-agnostic explanation method would be preferable.

A primary reason for focusing on local explanations stems from the sheer complexity of real-world problems. Figure 1 demonstrates how we can take advantage of local explanations to understand and solve the problem of interpretability in a decomposable manner. In this example, we trained a classifier using our fragility fracture dataset, with only two standardized variables – *L1L4_BMD* and *Total_BMD*. *L1L4_BMD* corresponds to the BMD levels in the lumbar spine, while *Total_BMD* refers to the overall BMD in the body. For demonstration purposes, we randomly sample 10 instances in the vicinity of each data instance in our dataset and highlight the predicted labels of those instances generated using a trained classifier. Panel 1 in Figure 1 is a visualization of all sampled instances – marked with a blue "X" if an instance is classified as negative -i.e., *without* fracture, and marked with a red dot if it is classified as positive – i.e., *with* a fracture. At first glance, it appears that the interaction between the two features is highly non-linear and messy, making it difficult to interpret on a global scale. Panel 2 is a subspace of the space shown in panel 1 – confining *L1L4_BMD* to specific value ranges, i.e., [-1.1, -0.92], and *Total_BMD* to [-0.2, -0.02]. Now, a clearer pattern that is recognizable by humans is starting to emerge – most positive instances have *L1L4_BMD* values in the range [-0.975, -0.950]. Lastly from panel 3, which is a subspace of panel 2 while confining *L1L4_BMD* to [-0.972, -0.95], and *Total_BMD* to [-0.17, -0.12], we can derive some inductive rules regarding the model. For instance, a green straight line in panel 3 is a local decision boundary that can be learned through local explanation models.

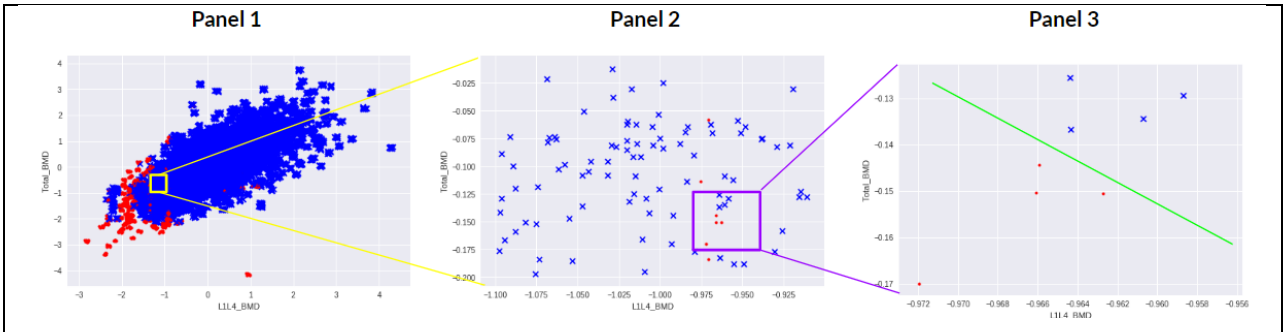

Figure 1. Comparison between instances on global and local scales

The local interpretable model-agnostic explanations (LIME) method (Ribeiro et al. 2016) is a widely accepted method for generating local, model-agnostic explanations. LIME uniformly samples training instances in the whole input space. Then, the instances are weighted proportional to the distance from the data instance to be explained. Finally, a local surrogate model – i.e., a *K*-Lasso model – is trained with weighted instances. However, it has several shortcomings, including the lack of objective evaluative criteria and diluted local faithfulness, which was recently recognized by Laugel et al. (2018). Accordingly, the latter proposed a new quantitative evaluation criterion for local explanations and improvements in the sampling

scheme for training instances to enhance local faithfulness with their local surrogate (LS) model. Nonetheless, this approach also has limitations and requires additional considerations when applied to real-world HPA problems. Thus, we attempt to improve upon both methods with our proposed ROLEX method in this study. Furthermore, we demonstrate the applicability and practical utility of our method while solving a real-world problem of fragility fracture prediction.

Proposed Method

Before explaining our proposed ROLEX method, we briefly describe general notation and common procedures of *local, model-agnostic explanation methods*. First, a binary black-box classifier ($f: \mathcal{X} \rightarrow Y$), e.g., deep neural networks or ensemble model, is trained with a finite number of data instances (x_i) in a p-dimensional input feature space $\mathcal{X}$. Following that, we make predictions for N instances ($x_i \in X \mid i \in \{1, 2, \dots, N\}$), which may or not may not be a subset of the training set. The objective of *local, model-agnostic explanation methods* is to develop a surrogate model ($s_{x_i} \in S_{\mathcal{X}}$) corresponding to each instance x_i that is inherently interpretable by humans and locally faithful to the classifier in the vicinity of that instance. Figure 2 shows an example demonstrating the process of generating local, model-agnostic explanations for a binary classifier f. In the figure, the blue region is where instances are negatively classified by f, whereas the orange region contains instances that are positively classified; one can see that the decision boundary is highly non-linear on a global scale. The local explanation models s_{x_1} and s_{x_2} are locally faithful to the black-box model f around x_1 and x_2, respectively. In other words, s_{x_1} and s_{x_2} effectively mimic the pattern of the classifier f in the neighboring region of x_1 and x_2 respectively. Furthermore, they are both linear models that can be easily interpreted by a human. In the context of the fragility fracture prediction problem, each data instance corresponds to a patient with p clinical or non-clinical characteristics that are likely to be useful in predicting future fragility fracture risks – e.g., BMD scores, the level of physical activity, and hormone levels. After predicting the fragility risk of N patients, i.e., N instances, using the black-box classifier f, we aim to generate independent explanations for the classifier's decision to classify patients as either having a fragility fracture risk or not having one. Finally, it should be emphasized that the local explanation models (s_{x_i}) are trained to primarily explain the black-box classifier (f), not to describe the phenomenon *per se*. Hence, regardless of the relationship between the black-box classifier and the phenomenon, the explanation model attempts to mimic the classifier. In some cases, the classifier may very well approximate the phenomenon of interest – in such cases, the local explanation models can be used to represent both the classifier and the phenomenon. However, the classifier may be inefficient in mimicking the complex relationships between variables in reality – i.e., an under-fitted prediction model. For instance, the classifier may be inappropriately trained due to the insufficient number of training samples or a highly skewed training dataset. Nonetheless, the explanation model can still be valuable in such scenarios. It can be used to point out where and why the classifier is underperforming to improve the classifier with human knowledge.

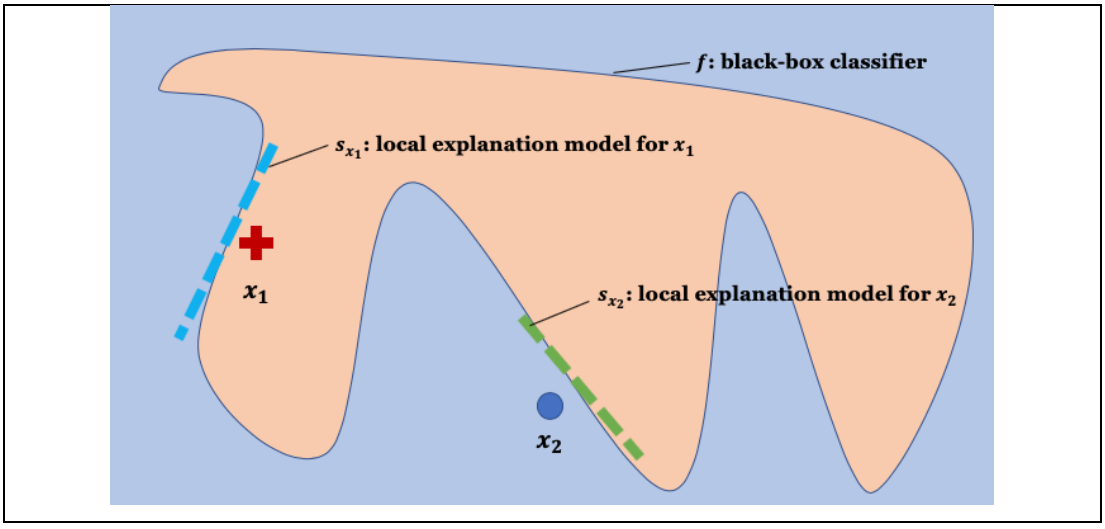

Figure 2. Illustration of local explanation models faithful to a black-box model

In this study, we propose the ROLEX method that improves over existing explanation methods – i.e., the LS method (Laugel et al. 2018), and the LIME method (Ribeiro et al. 2016). Also, we propose a new metric to more effectively evaluate the local faithfulness of explanation models – i.e., the *D-Fidelity* score. The ROLEX method has three phases – *pre-processing, training, and testing*. The pre-processing phase involves transforming the input feature set using the min-max scaling procedure. In the training phase, we combine the sampling procedure of the LS method and penalized model fitting procedure of LIME to train local explanation models. By integrating two established methods, we enhance the local faithfulness by restricting the influence of global features. We also improve the fitness of the surrogate model by penalizing for complexity and giving higher weight to closer points during the training phase. In the testing phase, we identify instances that are close to the decision boundary of f by sampling data points around them and ensuring that predictions made using the samples belong to both classes (i.e., positive and negative). The Local Fidelity scores (Laugel et al. 2018) are then computed for the identified decision-sensitive instances. As shown in Figure 1, training datasets in the medical domain are often highly non-linear and have skewed class distribution. In such cases, the local explanation models may have high fidelity scores despite fitting badly for data points far away from the decision boundary. Our D-Fidelity score addresses such issues by automatically considering only valid points that are close to the decision boundary. Figure 3 summarizes the ROLEX method, which is explained in detail in the following sub-sections.

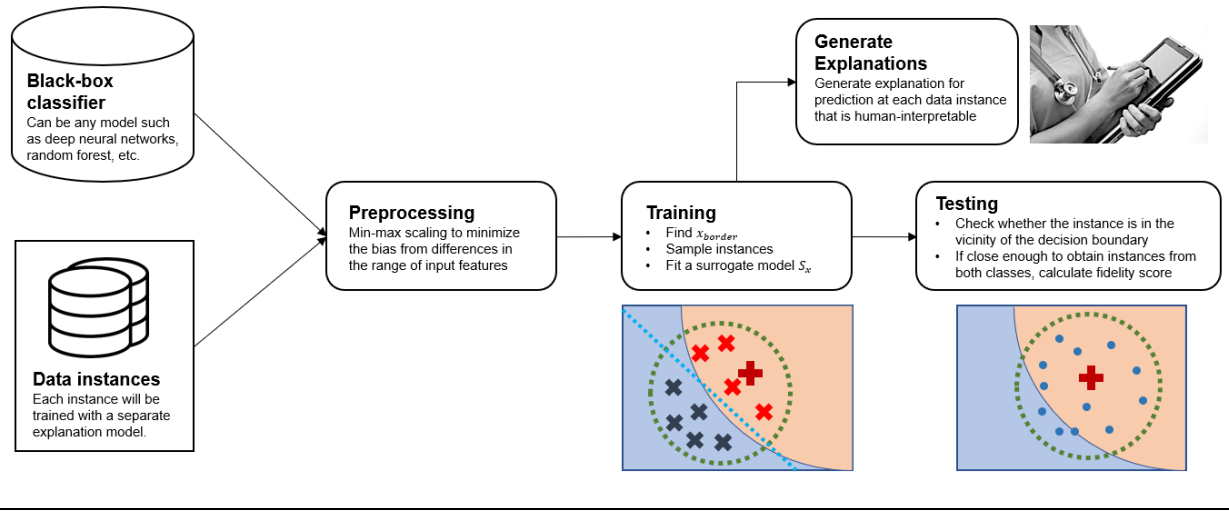

Figure 3. Flowchart of the proposed ROLEX method

Pre-processing Phase

In the pre-processing phase, we transform the inputs using the min-max scaling procedure given as:

$$X_{minmax} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

, where X_{min} and X_{max} correspond to the minimum and maximum values, respectively, of the input matrix X with p columns – i.e., with p features. The min-max scaling is carried out as a pre-processing step to remove bias due to differences in the range of input features and outliers. To understand the influence of different scales in input features, consider a hypothetical two-dimensional feature space consisting of two variables – a variable representing each *patient's height in centimeters* ($x^{(1)}$), and a binary variable indicating *whether a patient smokes or not* ($x^{(2)}$). In general, values of $x^{(1)}$ fall into a relatively small range of [0, 1], and values of $x^{(2)}$ may have a range of [150, 200]. Due to the larger range of $x^{(2)}$, it will have a larger influence than feature $x^{(1)}$ in the model fitting procedure for local explanation models. Since the local explanation methods generally use standard distance measures such as Euclidean distance, the resulting local explanation model (s_{x_i}) is likely to be biased if we do not scale the features. We explored different

scaling procedures, including normalization, standardization, and box-cox transformation, and found min-max scaling to have the best performance in terms of local faithfulness. Finally, it should be noted that the primary objective of employing a scaling scheme in our ROLEX method is to maximize *the local faithfulness of an explanation model* (s_{x_i}), not to enhance *the predictive performance of a prediction model (f)*.

Training Phase

The training phase of the ROLEX method has two steps: (1) identifying the local neighborhood for generating training instances, and (2) fitting the local explanation model (s_{x_i}) over locally sampled instances (Algorithm 1 below). Figure 4 is an illustration of the training phase of the ROLEX method for a single instance, denoted as x_1. The first step of identifying the local vicinity for generating training samples involves identifying a suitable region close to the decision boundary defined by the black-box class ifier for a given instance x_i. x_{border} is defined as a data point in the input feature space $\mathcal{X}$, which is closest to x_i with an oppositely predicted label. In Figure 4, $x_{border}$is the closest point that falls on the blue region that is closest to the instance x_1. In the fragility fracture prediction context, if x_i is predicted as having a fracture in the future, x_{border} should be predicted as not having fracture by f. The initial radius of the hypersphere (r_{init}) is arbitrarily set by the user (step 1-(i) of Algorithm 1). The radius is shrunk and expanded in the iterative sampling process for finding x_{border}. A sufficient number (n_{sample}) of instances are uniformly sampled in the hypersphere (step 1-(ii) of Algorithm 1) and if at least one instance with an opposite label is detected in the neighborhood, the radius is reduced by 50% (step 1-(iii) of Algorithm 1). If there is no instance with an opposite label among n_{sample} instances, the radius (r') is recorded (step 1-(iv) of Algorithm 1). Then instances are again sampled in the area between two hyperspheres, one having the converged radius (r') and the other two times the radius ($2r'$), (step 1-(v) of Algorithm 1). If there is at least one instance with an opposite label among the newly sampled instances, set one of them as x_{border}. Otherwise, we keep enlarging the hyperspheres and iteratively sample n_{sample} instances to find one (1-(vi) of Algorithm 1). Here, the distance between instances is measured using the Euclidean distance measure.

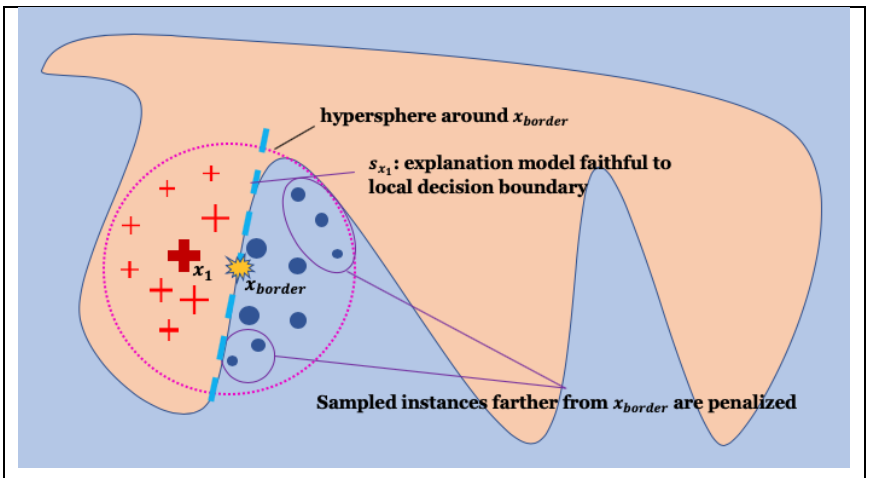

Figure 4. Training phase of ROLEX method for instance x_1

This is followed by uniformly sampling training instances, i.e., z_i's ($i = 1, \dots N_{train}$), around x_{border} up to a specific distance, i.e., sampling in a hypersphere around x_{border} with a pre-defined radius r_{train}. The radius of the hypersphere (r_{train}) can be set by the user (2-(i) of Algorithm 1). Then, a penalized logistic regression model s_x is trained over the sampled instances as a local explanation model for x, with residuals weighted based on the distance of samples from x_{border}. In Figure 4, one can see that the relative size of instances, which indicates the relative weight, varies based on the distance to x_1; the larger the distance, the smaller the weight. In the case of datasets with a large number of features, i.e., when p is large, the logistic regression model is usually penalized based on model complexity using the L-1 regularization scheme (Tibshirani 1996) to generate sparse explanations (step 2-(ii) of Algorithm 1). The resulting decision boundary of s_x is indicated with a dashed blue line in Figure 4.

Input: A binary classifier ($f: \mathcal{X} \rightarrow \{-1,1\}$) and data instances to be explained ($x \in \mathcal{X}$)

Initialize: $r_{init}, r_{train}, n_{sample}, N_{train}$

Algorithm:

1. Find the nearest instance x_{border}:
   i. Set $r = r_{init}$
   ii. Sample instances $(z \in Z \mid \|x - z\|_2 \leq r)_{n_{sample}}$.
   iii. If $\exists\, f(x) \neq f(z)\ \forall\, z \in Z$ set $r = r/2$ and go to ii. Else, set $r' = r$.
   iv. Set $a_l = r', a_u = 2r'$.
   v. Sample instances $(z' \in Z' \mid a_l \leq \|x - z'\|_2 \leq a_u)_{n_{sample}}$.
   vi. If $\exists\, f(x) \neq f(z')\ \forall\, z' \in Z'$ set $x_{border} = \underset{z'}{\operatorname{argmin}} \|x - z'\|_2$. Else set $a_l = a_u, a_u = a_u + r'$ and go to v.
2. Fit $s(x)$
   i. Generate training samples $(k \in K_x \mid \|k - x_{border}\|_2 \leq r_{train})_{N_{train}}$.
   ii. Model $s(x)$ as a regularized weighted logistic regression model $logit\ E(Y) = \alpha + \beta K_x$ by minimizing the function $\sum_{i=1}^{N_{train}} w_i (y_i - \hat{y}_i)^2 + \lambda \sum |\beta|$

Output: Explanation model s_x locally faithful to the black-box classifier f in the neighborhood $x \in \mathcal{X}$

Algorithm 1. Training local surrogate model

To summarize, the training process of the ROLEX method adapts both the sampling scheme of LS (Laugel et al. 2018) and the optimization scheme of LIME (Ribeiro et al. 2016). We find that combining these two result in enhanced local faithfulness of explanations, as described in the next section. In addition, we derived some practical considerations when applying the local explanations method with real-world datasets. The time complexity for the algorithm is $O(n^2) \leq O(.) \leq O(n^3)$ where n is the number of data instances to be explained. For each instance, the algorithm to find x_{border} - i.e., the first part of Algorithm 1 - will converge within $O(n^2)$. The worst-case scenario is when there are only instances of the opposite class in the feature space (excluding the given instance requiring explanation), and *n-1* iterations are required for instances in the majority class to establish the hypersphere radius (r'). The fitting procedure of an explanation model is linear in time for N_{train} and also depends on the regularization factor k. For a dataset with less than 100,000 instances, the proposed algorithm is expected to converge within a few hours. With our fragility fracture dataset having over 70 features and 600 test instances, the training algorithms for ROLEX and existing methods converged in less than 2 minutes with a standard single-core machine. Another important procedure is setting the initial value of parameters ($r_{init}, r_{train}, n_{sample}, N_{train}$), which can be determined by the user. While there are no systematic methods suggested in the previous works to determine such values, we have found that near-optimal values can be determined using standard hyper-parameter tuning procedures for standard machine learning models. The user can try out initial values based on his/her prior knowledge of the dataset as well as the problem domain. To further fine-tuning, a grid or random search algorithm for hyperparameter tuning (Bergstra and Bengio 2012) can be utilized.

Testing Phase

Instances with greater distance from the decision boundary are typically homogeneous – i.e., instances with the same predicted label. Such homogeneity in test instances will render many standard metrics to measure the performance of a prediction model meaningless. For instance, area under the curve (AUC), precision, and recall are not useful indicators of performance when the test dataset has only negative or positive samples. In fact, surrogate models learned with homogeneous sampled instances convey virtually no information – a classifier trained with only single-class instances is no less intelligent than an untrained one. If one attempts to evaluate such explanation models or use it to interpret a prediction model, he or she is likely to be headed to a wrong direction from the outset, leading to undesirable consequences. Such issues are not addressed in the LS method by Laugel et al. (2018), in which the local fidelity scores utilize standard

metrics for assessing the performance of surrogate models (s_x) in comparison to the original black-box classifier (f).

Therefore, in the testing phase, we first identify data instances closer to the decision boundary by sampling in their local neighborhood and checking if there are samples corresponding to both predicted class labels - e.g., both labels predicted as *having* a fracture and *not having* a fracture. Here, the "vicinity" or "closeness" is defined not as physical proximity, but more as the relative density of instances from both classes. Thus, if one can sample instances from both classes near an instance, it is an instance close to the decision boundary, regardless of the position in relation to other training/test instances. Thereafter, we calculate the decision-sensitive fidelity (*D-Fidelity*) scores for the instances near the decision boundary. When calculating *D-Fidelity* scores, instances far away from the decision boundary and consequently having homogeneous test instances are not considered. Hence, it is a better approximation of the local faithfulness of explanation models, compared to the original local fidelity score in Laugel et al. (2018). Algorithm 2 describes the steps for computing *D-Fidelity* scores.

In other words, instances whose *D-Fidelity scores* are not computable are instances for which local explanations models are likely to be invalid and less informative. In such cases, human decision-makers should not take the interpretations generated by local, model-agnostic explanation methods at facevalue. They are recommended to employ other explanation and interpretation methods that can be more suitable in such cases. For instance, the attention method (Choi et al. 2016) may be used to generate local explanations for recurrent models and localization methods such as Grad-CAM (Selvaraju et al. 2017) may be used for convolutional neural networks.

Input: Local explanation models $s(x) \in S(\mathcal{X})$, test instances ($x \in \mathcal{X}$), binary classifier ($f: \mathcal{X} \rightarrow \{-1,1\}$)

Initialize: r_{test}, n_{iter}, n_{sample}, N_{test}, *Acc (any standard evaluation metric for prediction models – e.g., AUC, precision, and recall)*

Algorithm:

1. Identify instances ($x' \in \mathcal{X}'$) close to decision boundary:

   For each x:

   i. Sample instances $(z \in Z \mid \|x - z\|_2 \leq r_{test})_{n_{sample}}$ for n_{iter} times.

   ii. If $\exists\, f(z_1) \neq f(z_2)\ \forall\, z_1, z_2 \in Z$, for each of n_{iter} samples, include x into a set of instances close to decision boundary ($x' \in \mathcal{X}'$).

2. Calculate D-fidelity score for each x':

   i. Sample $(z' \in Z' \mid \|x' - z'\|_2 \leq r_{test})_{N_{test}}$ in the vicinity of x'

   ii. Compute:

$$D\text{-}Fidelity(x', s_{x'}) = Acc_{z'_i \in Z'}(f(z'), s_x(z'))$$

Output: Decision-sensitive local fidelity scores $D\text{-}Fidelity(d', s_{d'})$.

Algorithm 2. Computing decision-sensitive local fidelity scores

Analysis

Data

The dataset used in this study was acquired from the Ansung cohort study, a prospective study started in 2001, supported by the National Genome Research Institute of the Korea Centers for Disease Control and Prevention (Kim et al. 2017). Data were collected using a large community-based epidemiological survey of a patient population of Korean men and women (40 to 69 years old). 5,018 participants completed a preliminary examination and survey in 2001. For our study, we selected a subset of participants with valid dual-energy X-ray absorptiometry (DXA) data indicating the presence or absence of a fracture. A set of features relevant for fragility fracture prediction was shortlisted by an experienced endocrinologist from a collection of over 3,000 indicators. The resulting dataset for analysis comprises of 2,227 instances and 75 attributes. The outcome is a binary variable indicating whether each individual developed a fracture during

the follow-up period of 10 years after starting the cohort study – i.e., between 2001 and 2011. The 10-year time window for assessing fracture risk conforms to the FRAX® guidelines for fracture risk identification. Out of 2,227 instances, 570 were reported to have a fragility fracture. The input attributes are classified into five categories, i.e., demographics, lifestyle, physical examination, medical records, and bone quality scores. Table 1 is the summary of the attributes and their categories, along with descriptions and some examples.

Table 1. Description of attributes

Category	Description	Examples
Demographics	Self-reported general information about the individual	Gender, age, income, education level
Lifestyle	Self-reported information about the habitual aspects of individual lives	Alcohol consumption, smoking, exercising
Physical examination	Records measured by professional medical examiners	Height, weight, hormone level, body fat percentage
Medical records	Previous medical records of individuals relevant to fragility fractures	Previous fracture occurrences, parental fracture occurrences
Bone quality scores	Scores indicating the quality of the bone structure	Trabecular bone score (TBS), bone mineral density (BMD)

As part of data pre-processing, instances with null values in outcomes and clinically identified predictors were excluded resulting in 2074 individuals. Then the dataset was partitioned into a training dataset with 1,451 instances and a testing dataset with 623 instances (i.e., 7:3 train-test ratio). Accuracy and AUC were chosen to evaluate the binary classifier. Here, accuracy is defined as the proportion of instances classified correctly by the model – i.e., the average of precision scores for two classes.

Fragility fracture predictions using machine learning

We compare the performance of six machine learning models for binary classification (bone fracture = 1, absence of bone fracture = 0) as shown in Table 2. The extreme gradient boosting (XGBoost) model shows the best accuracy and the categorical boosting (CatBoost) model shows the best AUC score. We conjecture that this is due to many categorical variables in our dataset – e.g., level of alcohol consumption. We consider the global model developed using CatBoost for further analysis as it shows the highest AUC score.

Table 2. Performance comparison of global modeling methods for bone fracture prediction

Method	Accuracy	AUC
Logistic regression	0.6292	0.6277
SVM	0.7431	0.4957
Feedforward Neural Networks	0.7207	0.5527
Random Forest	0.7271	0.5994
XGBoost	**0.7464**	0.6384
CatBoost	0.7384	**0.6522**

Trabecular bone score (TBS) of L1-L4 segments of the spine (*TBS_L1L4*), BMD of L1-L4 segments of the spine (*L1L4_BMD*), baseline bone mineral density score (*Total_BMD*), and Femoral Neckbone mineral density score (*Neck_BMD*) were identified to be the top 4 predictors for bone fracture by SVM and tree-based models. TBS is a complementary metric to BMD scores in measuring bone quality. While BMD considers the density of the bone, TBS examines bone microarchitecture that provides additional skeletal information (Shevroja et al. 2017). Considering a large number of collinear input features in the data, we fit a logistic ElasticNet regression (Zou and Hastie 2005) model to develop global explanations. The non-zero coefficients of the regression model are shown in Table 3. Nevertheless, such explanations are not sufficient in explaining individual-level differences among patients. In general, the user is interested in why a certain instance of interest relevant to him or her is classified as negative or positive, rather than the general trend.

Table 3. Coefficients of logistic ElasticNet regression model

Variables	Description	Estimate
TBS_L1L4	Vertebral TBS	-0.9551
Total_BMD	Total BMD	-1.7474
Neck_BMD	Neck BMD	-0.6858

AS4_age	Age	0.0038
as4_optrt	Whether the patient was treated for osteoporosis	0.0021
previousFx	Previous occurrence of fracture	0.3018
parental_Fx	Parental occurrence of fracture	-0.0027
as4_height	Height	-0.0082

Comparison of local explanation methods

We compared the performance of local explanation models trained using the ROLEX method with LIME and LS respectively, in terms of model fidelity. We considered the hypersphere radii in sampling for the training data (r_{train}) and test data (r_{test}) and set them to be equal to a $t * Max\_distance$ following the heuristic approach by (Laugel et al. 2018). t is a fraction assigned by the user and $Max\_distance$ is the maximum Euclidean distance between any two instances in the dataset. The value for t is set to be sufficiently small such that local explanation models can be defined for each instance, but large enough such that instances around the local decision boundary can be correctly classified. Optimal t can be determined using grid-search or using domain knowledge. Due to the min-max scaling procedure, the set of permissible t values for ROLEX is different from that of LIME and LS, as given in Table 4. The *D-Fidelity* scores of the models trained with different methods and different values of t are summarized in Table 4. Models fit using the ROLEX method have higher D-Fidelity scores than LS and LIME based models. The ROLEX method generates significantly more faithful local explanations than existing methods.

Table 4. Comparison of D-fidelity scores of local modeling methods

Method	Average D-fidelity (AUC)	Deviation	t
LIME	0.7632	0.1042	0.0001
	0.7233	0.0505	0.00001
	0.6174	0.0640	0.000001
LS	0.7499	0.1043	0.0001
	0.6119	0.1675	0.00001
	0.5419	0.1254	0.000001
ROLEX	0.8870	0.0369	0.01
	0.8322	0.0698	0.001
	0.8457	0.1001	0.0001

Local explanations and interpretations of fragility fracture predictions with the ROLEX method

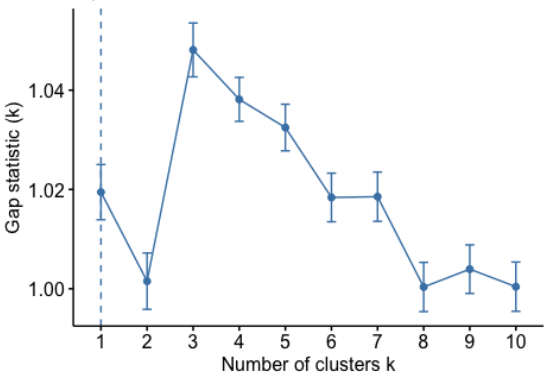

Figure 5. Gap statistic based on the number of clusters (*k*)

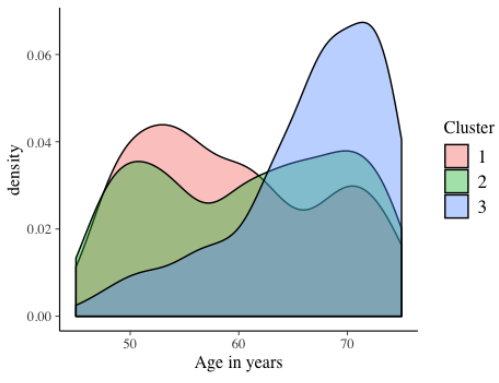

Figure 6. Age distribution in three identified clusters

To develop patient-level explanations for higher fragility fracture risks, we fit local explanation models using the ROLEX method for each of the 1451 training instances. From the min-max scaled coefficients B_X

of the local explanation models, the unscaled coefficients for the input features were derived as $\beta_X = \frac{\mathrm{B}_X}{X_{max} - X_{min}}$. The local input feature coefficients β_X indicate the heterogeneity across osteoporosis patients for fragility fracture risks. To explore latent patterns in fragility risk heterogeneity across patients, we performed k-means clustering over the four input predictors (*TBS_L1L4, L1L4_BMD, Total_BMD,* and *Neck_BMD*). Using the gap statistic (Tibshirani et al. 2001), 3 distinct clusters of patients were identified as shown in Figure 5. Cluster profiling revealed *Age* as the most distinctive patient attribute contributing to the heterogeneity in fragility fracture risks due to the four input predictors.

Figure 6 shows that the majority of patients in cluster 1 belong to the 50-60 years age group, while patients in cluster 3 are distinctly older than patients in cluster 1 and cluster 2 on average. To further understand how the coefficients of the top four predictors are different in the local models trained across instances belonging to the three clusters, we fit separate logistic regression models to the patient dataset in each cluster. Table 5 shows that the relationship between TBS for L1-L4 (*TBS_L1L4*) and fractures is significant for patients belonging to cluster 1 but not for cluster 2 and cluster 3. For a unit increase in *Neck_BMD*, the log-odds of bone fracture risk decreases by -2.8063 for Cluster 1 and by -1.5709 for cluster 2. That is, for a unit decrease in *Neck_BMD*, the odds of suffering a fracture in the future increases by 17 times for patients belonging to cluster 1, while for patients in cluster 2, the odds increase by 5 times. However, for cluster 3 patients, none of the four features can be used as a significant predictor of fragility fracture risk. Specific values of local coefficients allow more interesting and relevant interpretations. For cluster 1, model coefficients for three variables (i.e., *TBS_L1L4, Total_BMD, and Neck_BMD)* are reported to be significant which are also significant for the model fit over the entire dataset. In contrast, for cluster 3, whose constituents are much older, all three input variables are insignificant. Such patterns help us understand and articulate some simple decision rules such as: (1) *if you are younger than 60 years and have low TBS and BMD scores, you are highly likely to have a fragility fracture in the future*, (2) *if you are older than 60 years, the TBS and BMD scores are no longer reliable prognostic clinical markers of a future fracture risk*. The first might be somewhat trivial, given that TBS and BMD scores are established predictors for fragility fractures (Kanis et al. 2001). Nonetheless, the second rule provides a subtle yet important clinical insight - i.e., more attention and care should be taken to comprehensively predict and diagnose fractures among elderly individuals. Just looking at the standard clinical markers such as BMD and TBS might underestimate the fracture risks among individuals over 60 years of age and lead to hazardous, if not deadly, consequences. Interestingly, Stone et al. (2003) found corroborating evidence by analyzing data from a large cohort of women *aged 65 and older*. The results reveal that though most fractures are related to low BMD, the proportion of fractures directly attributable to low BMD is rather modest. They suggest that taking into account risk factors other than BMD is critical in designing effective prevention strategies.

Table 5. Coefficients of top four predictors across data from three clusters

Variables	Model for cluster 1	Model for cluster 2	Model for cluster 3	Model for entire training dataset
TBS_L1L4	**-0.8123* (0.3881)**	-0.0489 (0.1822)	-1.8728 (1.6521)	**-0.4131* (0.1045)**
L1L4_BMD	1.2793 (0.9999)	0.5207 (0.8109)	-1.6397 (1.0589)	-0.4674 (0.5292)
Total_BMD	**-2.7408* (0.6680)**	**-1.4749* (0.5601)**	-0.7185 (0.7546)	**-1.8739* (0.3432)**
Neck_BMD	**-2.8063* (0.6938)**	**-1.5709* (0.6255)**	-0.5201 (0.8047)	**-1.8560* (0.3735)**
* indicates p-value < 0.05, values inside the parenthesis are standard deviations of the coefficients				

Table 6. Local explanation model coefficients (top-5) for three random patients

Patient #30		Patient #51		Patient #492	
Variables	**Coefficient**	**Variables**	**Coefficient**	**Variables**	**Coefficient**
Weight	-0.0445	*HDL*	0.0427	*Weight*	-0.0412
HDL	0.0283	*Weight*	-0.0422	*HDL*	0.0324
Oa11	0.0259	*Body Fat*	0.0267	*Body Fat*	0.0274
KADL	-0.0226	*Height*	0.0208	*Total BMD*	-0.0179

TBS_L1L4	-0.0169	*Education*	-0.0172	*KADL*	-0.0165

Table 6 lists the top five coefficients of the explanation models generated by the ROLEX for three randomly chosen patients (patient #30, #51, and #492). Even though all these patients have -1 prediction labels (i.e., minimum risk of fragility fracture), their corresponding explanation models have different significant predictor coefficients. We observe from Table 6 that body weight (*Weight*), and cholesterol level (*HDL*) are important predictors for all three patients. While the patient's body weight is negatively correlated with fractures, their cholesterol level is positively correlated. Body fat percentage (*Body Fat*) is positively related to fracture risk for patients #51 and #492, whereas daily activity level (*KADL*) is negatively related to fracture risk for patients #30 and #492. However, there are some heterogeneous associative patterns across the three patients as well. For instance, for patient #30, arthralgia score (*Oa11*) is positively correlated with fractures. This follows widely accepted clinical knowledge since arthralgias, described as pain or stiffness in the joints, is positively related to low bone quality. For patient #51, the education level (*Edu*) is negatively correlated with bone fractures. One possible justification for this pattern that requires future validation is that individuals with low education levels are commonly associated with blue-collar jobs that entail strenuous physical activity. Long term strenuous activity is related to incremental allostatic load for the body often leading to bone loss. Finally, for patient #30 TBS score (*TBS_L1L4*) is a significant predictor of fracture risk, while on the other hand, BMD score (*Total_BMD*) is a significant fracture risk predictor for patient #492. Past studies report either BMD scores or TBS scores as independent indicators of bone quality (Shevroja et al. 2017). The example presented in Table 6 validates existing clinical understanding that both TBS and BMD are important and selective indicators of fracture risk for different patient groups. Although both predictors were significant in the global model (Table 3), we show that the heterogeneity in input effects can be identified only from the local explanation models generated using the ROLEX method.

Finally, it should be noted that the explanations produced by ROLEX differ from coefficient estimates or decision rules obtained from inherently interpretable prediction models such as logistic regression or decision trees. In prior sections, we emphasized that prediction models and explanation models play different roles in *local, model-agnostic explanation methods;* the explanation model attempts to interpret the prediction model. Furthermore, the ROLEX method can be applied to various complex prediction models such as deep neural networks and ensemble models (e.g., XGBoost, RandomForest, etc.), which do not have inbuilt interpretation mechanisms.

Discussions

In this study, we proposed the ROLEX method and demonstrated its ability to solve a real-world problem of providing patient-level explanations for bone fragility fracture predictions. The ROLEX method shows stable as well as superior performance in terms of local fidelity. It has the potential to foster reliable and trustworthy explanations, improving patient safety and user acceptance of complex HPA applications. To our knowledge, this is the first study that provides in-depth insights into the value of explanations in HPA and demonstrates a design artifact for clinical intelligence with a focus on transparency.

Practical Implications

Trust strongly influences the usability of IT artifacts, and reliable explanations play an important role in building trust (Gefen et al. 2003; Gregor and Benbasat 1999). Physicians can build trust in the model by interpreting local explanation model coefficients and their clinical relevance. Furthermore, physicians can derive new insights into fracture risks by exploring the individual explanation models for heterogeneous, such as contrasting patterns among different age groups as demonstrated in the previous section. Such individual or between-group differences are difficult, if not impossible, to unravel by global explanations. Moreover, new clinical insights gleaned through local explanations can contribute to patient safety and well-being. Physicians can improve diagnosis and patient care with the expanded knowledge base. Finally, recent changes in regulations and governmental policies call for compelling explanations in practice. In 2016, the White House recognized challenges in big data, e.g., incompleteness in input data and poorly designed systems, emphasizing transparency and accountability in algorithmic-based decision making (Executive Office of the President 2016). Further, starting May 2018, the European Union requires algorithms to provide explanations, i.e., providing the user with the "right to explanation," when the

predictions affect users at a significant level with the GDPR (European Parliament and Council of the European Union 2016).

Theoretical implications

Lin et al. (2017) discuss how solving healthcare problems with an HPA system is relevant to and can contribute to IS discipline, especially in design science research. Extending their discussions, we argue that our ROLEX method to explain results from such systems is relevant for many research streams in IS, including *healthcare IS*, *big data and predictive analytics*, and *design science*. First, our ROLEX method aims to improve patient safety and user acceptance, which are highly critical in the healthcare context (Ahmad et al. 2018). It has been established that trust plays a critical role in the user's acceptance of technology in various contexts, and we believe that it plays a more important role in the context of healthcare as users are generally more reluctant to change (Fichman et al. 2011). Based on our work, future studies can examine whether robust, local explanations generated by the ROLEX method actually improve the user's perceptions toward an IT artifact. Secondly, explanations with the help of domain experts and the reverse-designing process can support the discovery of new knowledge regarding the phenomenon of interest (Doshi-Velez and Kim 2017). In the context of big data analytics, this can be one way to exploit the synergistic relationship between big data and theory (Rai 2016). The iterative, granular investigation carried out by implementing the prediction-explanation loop can significantly expedite this knowledge discovery process. In the analysis section, we showed how our ROLEX method can be used to understand heterogeneity across patient groups and homogeneity within each group. Such understanding may help experts discover new patterns inductively, and in turn, they can enhance existing knowledge, or develop new theories based on the patterns.

Limitations and future research

Our study has some limitations. There is a trade-off between the quality of training samples at the cost of computational time which is related to the initial value of the hypersphere radius, $\boldsymbol{r_{init}}$, in the surrogate modeling procedure. The lower the value of $\boldsymbol{r_{init}}$, the samples generated for training the local surrogate model will be more faithful to the instance to be explained, while taking more time to determine $\boldsymbol{x_{border}}$. Finding the optimal point considering both costs can be a potentially fruitful research direction**.** Another limitation of our method is commonly shared by local explanation models for producing multifaceted interpretation mechanisms. Global linear models have a single set of coefficients for generating interpretations, while in case of local modeling, each instance has its own set of models and coefficients, requiring the end-user to inspect each individually, or sample representative instances for comparative inference, or use a grouping strategy for interpreting multiple instances with a similar pattern of model coefficients. Similar to the clustering approach used in this study, future work can focus on the systematic exploration of instance-level coefficients and modeling heterogeneity.

Conclusions

In this study, we introduce a new method for generating local explanations for safe and user-centric healthcare predictive analytics and demonstrate its application in solving the bone fracture prediction problem. Our key contributions are two-fold: (1) proposing a novel, model-agnostic local explanation method by improving state-of-the-art methods and suggesting a new evaluation framework, and (2) advancing the theoretical discussion on HPA and the role of explanations in modern artificial intelligence systems in light of IS research. With our proposed method, we are not only able to interpret the predictions of the global model at the instance level of a single patient, but also, able to provide enlightening explanations for assessing fracture risks across heterogeneous patient groups. Further, we show that developing explanation methods for reliable and user-centric systems can contribute to the IS knowledge base, particularly in *healthcare IS, big data analytics, and design science*.

References

Ahmad, M. A., Teredesai, A., and Eckert, C. 2018. "Interpretable Machine Learning in Healthcare," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, New York, NY: IEEE, June, pp. 447–447.

Atkinson, E. J., Therneau, T. M., Melton, L. J., Camp, J. J., Achenbach, S. J., Amin, S., and Khosla, S. 2012. "Assessing Fracture Risk Using Gradient Boosting Machine (GBM) Models," *Journal of Bone and Mineral Research* (27:6), pp. 1397–1404.

Bahdanau, D., Cho, K., and Bengio, Y. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate," *International Conference on Learning Representations*.

Bardhan, I., Oh, J. (Cath), Zheng, Z. (Eric), and Kirksey, K. 2015. "Predictive Analytics for Readmission of Patients with Congestive Heart Failure," *Information Systems Research* (26:1), pp. 19–39.

Bergstra, J., and Bengio, Y. 2012. "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research* (13), pp. 281–305.

Black, D. M., Thompson, D. E., Bauer, D. C., Ensrud, K., Musliner, T., Hochberg, M. C., Nevitt, M. C., Suryawanshi, S., and Cummings, S. R. 2000. "Fracture Risk Reduction with Alendronate in Women with Osteoporosis: The Fracture Intervention Trial," *The Journal of Clinical Endocrinology & Metabolism* (85:11), pp. 4118–4124.

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. 2016. "Interpretable Deep Models for ICU Outcome Prediction," *AMIA Annual Symposium Proceedings* (2016), pp. 371–380.

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165–1188.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. 2016. "RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), Barcelona, Spain: Curran Associates, Inc., pp. 3504–3512.

Doshi-Velez, F., and Kim, B. 2017. "Towards A Rigorous Science of Interpretable Machine Learning," *ArXiv:1702.08608 [Cs, Stat]*.

European Parliament and Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council*, Official Journal of the European Union.

Executive Office of the President. 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, The White House.

Fichman, R. G., Kohli, R., and Krishnan, R. (eds.). 2011. "The Role of Information Systems in Healthcare: Current Research and Future Trends," *Information Systems Research* (22:3), pp. 419–428.

Forgetta, V., Keller-Baruch, J., Forest, M., Durand, A., Bhatnagar, S., Kemp, J., Morris, J. A., Kanis, J. A., Kiel, D. P., McCloskey, E. V., Rivadeneira, F., Johannson, H., Harvey, N., Cooper, C., Evans, D. M., Pineau, J., Leslie, W. D., Greenwood, C. M., and Richards, J. B. 2018. "Machine Learning to Predict Osteoporotic Fracture Risk from Genotypes," *BioRxiv*.

Gefen, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly* (27:1), p. 51.

Gregor, S., and Benbasat, I. 1999. "Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice," *MIS Quarterly* (23:4), p. 497.

Iliou, T., Anagnostopoulos, C.-N., and Anastassopoulos, G. 2014. "Osteoporosis Detection Using Machine Learning Techniques and Feature Selection," *International Journal on Artificial Intelligence Tools* (23:05), p. 1450014.

International Osteoporosis Foundation. 2018. "Broken Bones, Broken Lives - the Fragility Fracture Crisis in Six European Countries," , October 19.

Kanis, J. A., Johnell, O., Oden, A., De Laet, C., Dawson, A., and Jonsson, B. 2001. "Ten Year Probabilities of Osteoporotic Fractures According to BMD and Diagnostic Thresholds," *Osteoporosis International* (12:12), pp. 989–995.

Kanis, J. A., Oden, A., Johansson, H., Borgström, F., Ström, O., and McCloskey, E. 2009. "FRAX® and Its Applications to Clinical Practice," *Bone* (44:5), pp. 734–743.

Kim, Y., Han, B.-G., and the KoGES group. 2017. "Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium," *International Journal of Epidemiology* (46:2), pp. e20–e20.

Laney, D. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety," *Application Delivery Strategies*.

Laugel, T., Renard, X., Lesot, M.-J., Marsala, C., and Detyniecki, M. 2018. "Defining Locality for Surrogates in Post-Hoc Interpretablity," in *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, Stockholm, Sweden, June 19.

Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., and Yang, H.-J. 2017. "Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach," *MIS Quarterly* (41:2), pp. 473–495.

Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., and O'Connor, P. J. 2014. "A Machine Learning Approach to Improving Dynamic Decision Making," *Information Systems Research* (25:2), pp. 239–263.

Molnar, C. 2019. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*.

National Clinical Guideline Centre. 2012. *Osteoporosis: Fragility Fracture Risk*, (Vol. NICE Clinical Guidelines), Royal College of Physicians (UK).

NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy. 2001. "Osteoporosis Prevention, Diagnosis, and Therapy," *JAMA: The Journal of the American Medical Association* (285:6), pp. 785–795.

Rai, A. 2016. "Synergies Between Big Data and Theory," *MIS Quarterly* (40:2), pp. 3–10.

Rai, A. 2017. "Editor's Comments: Diversity of Design Science Research," *MIS Quarterly* (41:1), pp. 3–18.

Reginster, J.-Y., and Burlet, N. 2006. "Osteoporosis: A Still Increasing Prevalence," *Bone* (38:2), pp. 4–9.

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, San Francisco, California, USA: ACM, February 16.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 2017. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, October, pp. 618–626.

Shevroja, E., Lamy, O., Kohlmeier, L., Koromani, F., Rivadeneira, F., and Hans, D. 2017. "Use of Trabecular Bone Score (TBS) as a Complementary Approach to Dual-Energy X-Ray Absorptiometry (DXA) for Fracture Risk Assessment in Clinical Practice," *Journal of Clinical Densitometry* (20:3), pp. 334–345.

Silverman, S. L., and Calderon, A. D. 2010. "The Utility and Limitations of FRAX: A US Perspective," *Current Osteoporosis Reports* (8:4), pp. 192–197.

Simon, H. 1996. *The Sciences of the Artificial*, MIT press.

Stone, K. L., Seeley, D. G., Lui, L.-Y., Cauley, J. A., Ensrud, K., Browner, W. S., Nevitt, M. C., and Cummings, S. R. 2003. "BMD at Multiple Sites and Risk of Fracture of Multiple Types: Long-Term Results From the Study of Osteoporotic Fractures," *Journal of Bone and Mineral Research* (18:11), pp. 1947–1954.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)* (58:1), pp. 267–288.

Tibshirani, R., Walther, G., and Hastie, T. 2001. "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (63:2), pp. 411–423.

Tsuda, T. 2017. "Epidemiology of Fragility Fractures and Fall Prevention in the Elderly: A Systematic Review of the Literature," *Current Orthopaedic Practice* (28:6), pp. 580–585.

US National Library of Medicine. 2008. *Bone Density*.

Zou, H., and Hastie, T. 2005. "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (67:2), pp. 301–320.