medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Interpretable machine learning prediction of all-cause mortality

Wei Qiu[1], Hugh Chen[1], Ayse Berceste Dincer[1], and Su-In Lee[1,*]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[*]Corresponding: suinlee@cs.washington.edu

Abstract

Explainable artificial intelligence provides an opportunity to improve prediction accuracy over standard linear models using "black box" machine learning (ML) models while still revealing insights into a complex outcome such as all-cause mortality. We propose the IMPACT (Interpretable Machine learning Prediction of All-Cause morTality) framework that implements and explains complex, non-linear ML models in epidemiological research, by combining a tree ensemble mortality prediction model and an explainability method. We use 133 variables from NHANES 1999–2014 datasets (number of samples: $n = 47,261$) to predict all-cause mortality. To explain our model, we extract local (i.e., per-sample) explanations to verify well-studied mortality risk factors, and make new discoveries. We present major factors for predicting x-year mortality ($x = 1, 3, 5$) across different age groups and their individualized impact on mortality prediction. Moreover, we highlight interactions between risk factors associated with mortality prediction, which leads to findings that linear models do not reveal. We demonstrate that compared with traditional linear models, tree-based models have unique strengths such as: (1) improving prediction power, (2) making no distribution assumptions, (3) capturing non-linear relationships and important thresholds, (4) identifying feature interactions, and (5) detecting different non-linear relationships between models. Given the popularity of complex ML models in prognostic research, combining these models with explainability methods has implications for further applications of ML in medical fields. To our knowledge, this is the first study that combines complex ML models and state-of-the-art feature attributions to explain mortality prediction, which enables us to achieve higher prediction accuracy and gain new insights into the effect of risk factors on mortality.

1 Introduction

Machine learning (ML) has spurred significant advances in healthcare, especially in the vital areas of diagnosis and prognosis. Notable diagnostic studies include diabetic retinopathy detection in ophthalmology images [1], lung cancer classification from histopathology images [2] and skin cancer classification [3]. Many of these

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

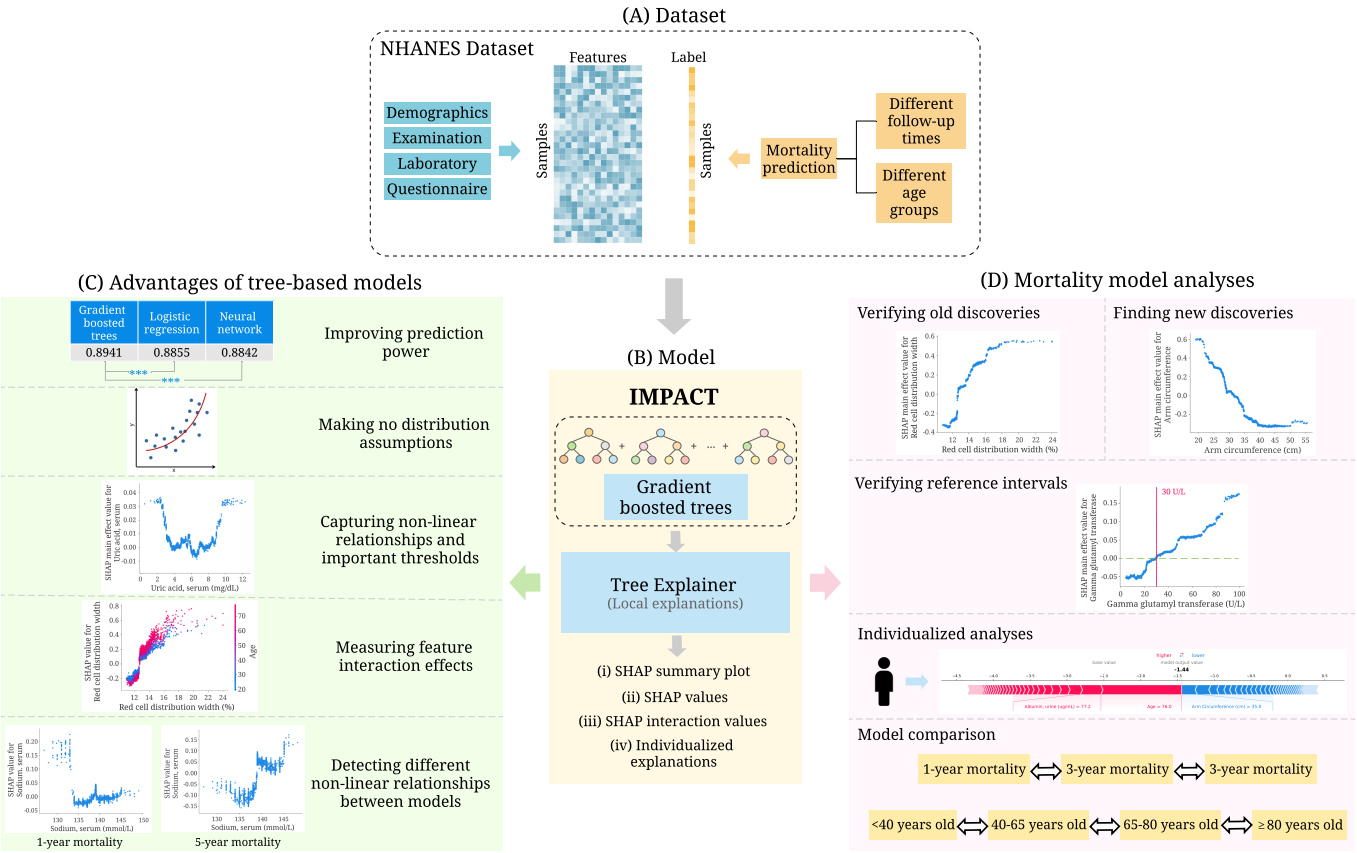


Figure 1: **Overview of the IMPACT model and analyses.** (A) We use the NHANES (1999-2014) dataset which includes 133 variables and 47,261 samples. The variables can be categorized into four groups: demographics, examination, laboratory and questionnaire. We train the model using different follow-up times and different age groups. (B) IMPACT combines tree-based models with an explainability method. Specifically, IMPACT (1) trains tree-based models are trained for mortality prediction using NHANES dataset (2) uses TreeExplainer to provide local explanations for our models. (C) We illustrate the advantages of interpretable tree-based model compared to traditional linear model in epidemiological studies. (D) We further analyze all mortality models and demonstrate the effectiveness of IMPACT to verify existing findings, identify new discoveries, verify reference intervals, obtain individualized explanations, and compare models using different follow-up times and age groups.

studies report accuracy comparable to or exceeding trained physicians and pathologists. Prognostic research examples include cardiovascular risk prediction using routine clinical data [4] and hospital readmission risk prediction [5]. Despite this progress, the primary limitation of complex ML for healthcare applications is that they are viewed as a "black box", which refers to the lack of *explainability* (i.e., we do not know why the model makes a certain prediction). In many cases, the patterns a model discovers are even more important than its predictive accuracy. This is especially true in prognostic research, which aims to identify important variables in order to make interventions, or to detect new risk factors to investigate further. The goal of applying ML to healthcare problems if often not only to achieve high accuracy, but to *explain* the prediction.

Identifying features associated with mortality is an important prognostic issue in public health and crucial for clinical decision making. Epidemiological studies that explore the relationship between measurements –

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

obtained by questionnaires, physical assessments and biological samples – and mortality identify risk factors provide new hypotheses about health determinants. Most of prior studies focus on the association of one risk factor with mortality at a time [6–8], using linear models; only a few papers have investigated multiple risk factors with linear models concurrently [9, 10]. Even though linear models are transparent, advanced ML models can achieve higher predictive accuracy compared to linear models because they capture interactions among variables. However, the few studies that use non-linear models for mortality prediction [11] focus only on their predictive power, and fail to explain the "black box" model to make new discoveries, essential for drawing epidemiological conclusions.

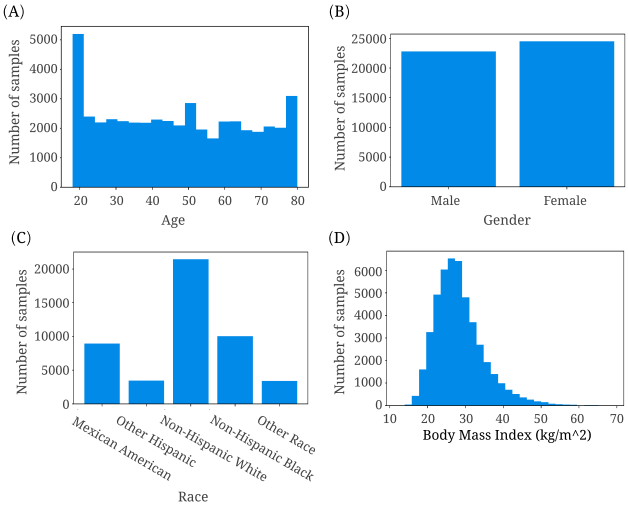
In this paper, we present the Interpretable Machine learning Prediction of All-Cause morTality (IMPACT; Figure 1) framework, which uses a complex, high-performance tree-based ML model and state-of-the-art explanation method to predict all-cause mortality. We adopt the tree-based models because: (1) Tree-based models outperform both linear models and neural networks across all mortality prediction tasks and age categories we consider. (2) Unlike linear models, tree-based approaches do not use a fixed parametric form and can capture non-linear relationships and feature interactions, which are crucial for epidemiological studies. (3) Recent advances have provided methods to rigorously explain tree-based models. TreeExplainer [12], a local explanation (i.e., feature attribution) method for trees, explains not only global decisions, but the impact of each variable for an individual prediction. It can identify important variables, explain model decisions and reveal non-linear relationships and feature interactions. Several medical studies have used TreeExplainer to extract valuable discoveries from "black box", tree-based models [13, 14] to generate new hypotheses. Importantly, in addition to providing accurate prediction scores, interpretable tree-based models can support clinical decision making by revealing individualized risk factors of mortality.

We applied IMPACT to the National Health and Nutrition Examination Survey (NHANES) (1999-2014) dataset. To our knowledge, this is the first study that uses interpretable complex ML models to do a systematic and integrated study of the associations between a large number of variables (133 variables) and all-cause mortality. We trained the model using different follow-up times (1-year mortality, 3-year mortality and 5-year mortality) and different age groups ($<$40 years, 40-65, 65-80 and $\geq$80). We discuss the advantages of tree-based models compared to linear models and neural networks (Section 3.1), analyze the discoveries obtained by 5-year mortality prediction models (Section 3.2), and compare the models for different follow-up times and age groups (Section 3.3) to identify the changes of importance and effect of variables on mortality.

Our contribution is summarized as follows:

- To our knowledge, IMPACT is the first attempt to combine complex ML models and state-of-the-art feature attribution method to explain all-cause mortality.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



(E)

	Number of samples	Number of alive samples	Number of deceased samples
1-year mortality	47,261	46,684	577
3-year mortality	41,434	39,537	1,897
5-year mortality	35,854	32,780	3,074
Age < 40	14,227	14,040	187
40 ≤ Age < 65	12,871	12,153	718
65 ≤ Age < 80	7,457	5,864	1,593
Age ≥ 80	1,299	723	576

Figure 2: (A)-(D) Histograms of age, gender, race, and body mass index in the NHANES dataset. (E) The sample size and number of living and deceased samples for different follow-up times and different age groups. For different age groups, the follow-up time is set to 5 years.

- We train highly accurate IMPACT models using a large number of variables, and do a systematic study of mortality.
- We illustrate the advantages of tree-based models using findings from IMPACT: (1) improving prediction power, (2) making no distribution assumptions, (3) capturing non-linear relationships and important thresholds (e.g., uric acid), (4) measuring feature interaction effects (e.g. red cell distribution width v.s. age), and (5) detecting different relationships among models (e.g., serum sodium).
- We identify new discoveries from IMPACT models including little-studied risk factors (e.g., platelet count), interaction effects (e.g., serum chloride v.s. age and gender) and reference intervals (e.g.,blood lead).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2 Methods

2.1 Data collection and processing

The National Health and Nutrition Examination Survey (NHANES) from the National Center for Health Statistics (NCHS) [1] conducts interviews and physical examinations to assess the health and nutrition data for all ages in the United States. The interviews include demographic, socioeconomic, dietary, and health-related questions. The examinations include medical, dental, physiological measurements, and laboratory tests administered by highly trained medical personnel. Since 1999, data were collected and released at 2-year intervals. Each year NHANES examines a nationally representative sample of roughly 5,000 individuals across the Unites States. In this study, we include NHANES data sampled between 1999 and 2014. All-cause mortality is ascertained by a linked NHANES mortality file that provides follow-up mortality data from the date of survey participation through December 31, 2015.

Our study includes samples with known mortality status who participated in NHANES 1999-2014 ($n = 47,261$). We include all demographic, laboratory, examination, and questionnaire features that could be automatically matched across different NHANES cycles. We exclude variables that are missing for more than 50% of the participants and highly correlated features with correlations greater than 0.98; after filtering, 133 features remain. We impute missing data using MissForest [15], a nonparametric random forest-based multiple imputation method for mixed-type data, with seven iterations. Figure 2A-D show distributions of a few basic features. We predict all-cause mortality for two broad categories: (1) follow-up times of 1-year, 3-year, and 5-year and (2) age groups of <40, 40-65, 65-80, and ≥80 years old. For different follow-up times, we remove samples with unconfirmed mortality status. For different age groups, we predict 5-year mortality. The sample size and number of living and deceased samples are shown in Figure 2E.

2.2 Predictive modeling

To model mortality, we use gradient boosted trees (GBTs) which achieve state-of-art performance on many tabular datasets [16, 17]. GBTs are nonparametric methods composed of iteratively trained decision trees. The final ensemble of trees captures non-linearity and interactions between predictors. We use the implementation XGBoost [18] with a learning rate set to 0.002 and 10,000 trees of max depth 3. For comparison, we also train logistic regression models and deep neural networks. For logistic regression, we use L2 regularization. For neural networks, we use a single layer with 1,000 nodes, and max iteration set to 1,000. The hyperparameters for logistic regression and neural networks are chosen by GridSearch and cross validation. The dataset is randomly divided into training (80%) and testing (20%) sets. Models' performance is measured with the area

[1] http://www.cdc.gov/nchs/nhanes.htm

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

under the receiver operator characteristic curve (AUROC). We bootstrap the test set to assess the statistical significance of the difference in AUROC for pairs of models. Specifically, we resample with replacement from the test set 1,000 times and compare the models' performance on resampled test sets. We report a p-value which is the percentage of time that logistic regression or the neural network's performance is better than or equal to gradient boosted trees, divided by the number of resampled test sets. All models are built using the Scikit-learn package in Python 3.7.

2.3 Model interpretation

To explain the GBT models, we utilize TreeExplainer [12], which provides a local explanation of the impact of input features on individual predictions. Specifically, TreeExplainer calculates exact SHAP [19] (SHapley Additive exPlanations) values, which guarantee a set of desirable theoretical properties. First, SHAP values are *additive*. They sum to the model's output, i.e., the log-odds for GBTs. Second, they are *consistent*, which means features that are unambiguously more important are guaranteed to have a higher SHAP value. Therefore, SHAP values are consistent and accurate calculations of each feature's contribution to the model's prediction. In our study, higher SHAP values imply higher mortality risk. TreeExplainer also extends local explanations to directly capture feature interactions and explains global model structure based on many local explanations. By showing the impact of each variable and the impact of interactions between variables for sample-specific explanations, we can obtain a comprehensive understanding about why the model made a specific prediction.

3 Results

3.1 Advantages of tree-based models

3.1.1 Tree-based models make more accurate predictions than linear models and neural networks

Linear models are commonly used because their coefficients provide a global explanation, which indicate each feature's contribution to the model's prediction [20]. However, for many datasets, more expressive models, such as tree-based models and neural networks, can achieve higher predictive accuracy by learning non-linear relationships between independent and dependent variables. The advantages of computational efficiency and high accuracy have made tree-based models the most popular non-linear model class, especially for tabular-style datasets where features are individually meaningful (as opposed to images/text/time series data which are often dominated by neural networks) [21]. Tree-based models — such as random forests and

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

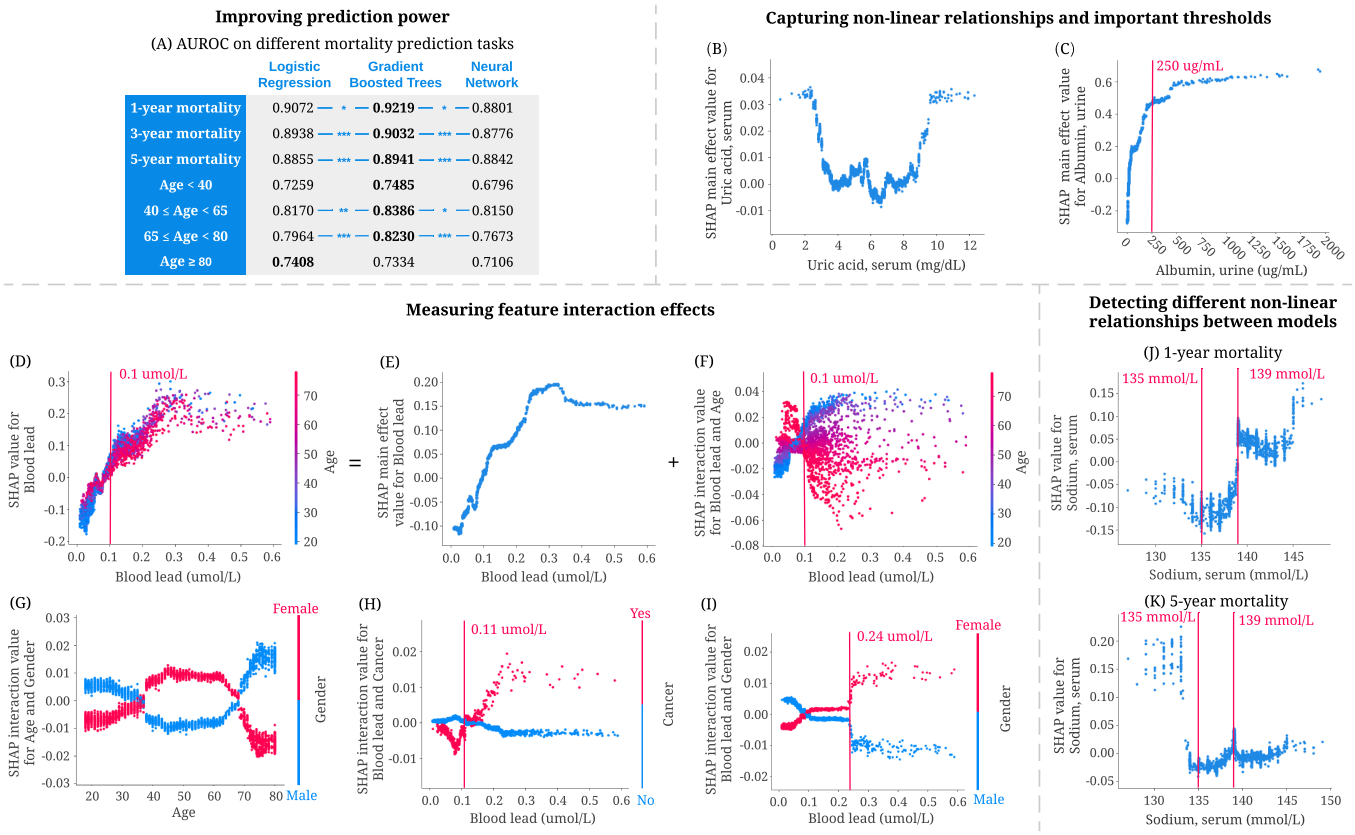


Figure 3: **Advantages of tree-based models for mortality prediction.** (A) The area under the ROC curve (AUROC) of gradient boosted tree models outperforms both linear models and neural networks for six of our prediction models. (Section 3.1.1) .($***$) represents a P-value < 0.001, ($**$) represents a P-value < 0.01, and ($*$) represents a P-value < 0.05. P values are computed using bootstrap resampling over the tested time points while measuring the difference in area between the curves. (B,C) Tree-based models can capture non-linear relationships and important thresholds (Section 3.1.3). (B) The main effect of uric acid on 5-year mortality. Higher SHAP value leads to higher mortality risk (C) The main effect of urine albumin on 5-year mortality. (D–I) Tree-based models can measure feature interaction effects (Section 3.1.4). (D) SHAP value for blood lead level in the 5-year mortality model. Each dot corresponds to an individual. (E) We can use SHAP interaction values to remove the interaction effect of age from the model and obtain the main effect of blood lead on 5-year mortality. (F) Plotting just the interaction effect of blood lead with age shows how the effect of blood lead on mortality risk varies with age. (G) The SHAP interaction value of age vs. gender in the 5-year mortality model. (H) The SHAP interaction value of blood lead vs. cancer in the 5-year mortality model. (I) The SHAP interaction value of blood lead vs. gender in 5-year mortality model. (J,K) Tree-based models can highlight different non-linear relationship between models. (Section 3.1.5) (J) The SHAP value of serum sodium in the 1-year mortality model. (K) The SHAP value of serum sodium in the 5-year mortality model.

gradient boosted trees — achieve state-of-the-art performance in many domains [22–25]. We observe the same trend when predicting mortality using our NHANES data set: tree-based models outperform both linear models and neural networks across all mortality prediction tasks and age categories we consider (Figure 3A). The superior prediction performance of tree models indicates that we can capture signals relevant to mortality, which alternative approaches could not.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

3.1.2 Tree-based models make minimal assumptions about the data distribution

Several principle assumptions are associated with linear models: linearity, independence, normality, homoscedasticity, and little or no multicollinearity. Of particular note, the linearity assumption states that the relationship between independent and dependent variables is linear. To satisfy this assumption, scientists often manually transform non-linear variables prior to fitting a model (e.g., log-transformation, discretization of continuous variables, etc.). Other assumptions also restrict the features linear models can use. In comparison, nonparametric methods like tree-based models make minimal assumptions about the data distribution and need no data transformations.

For instance, Figure 3D–E show the SHAP values and main effects of blood lead for a 5-year mortality prediction model. SHAP values indicate the contribution of each feature to model prediction where higher SHAP values indicate higher model prediction and vice versa. The main effect is the impact of a single feature, computed by subtracting the impact of all other features from the SHAP value (e.g., interaction effect). The figure shows a positive relationship between blood lead and 5-year mortality risk, which matches the conclusion of previous studies that used linear models [26–28]. In these papers, researchers first discretized blood lead using different thresholds and observed that individuals with blood lead levels higher than the threshold had increased all-cause mortality risk compared to those with lower blood lead levels. In stark contrast, tree-based models can capture non-linear relationships, which eliminates the need to manually transform the variables. Furthermore, using local explanations, we can identify the importance of blood lead for any continuous value, rather than for discrete blood level categories, which makes it possible to get *personalized* explanations that are explicit to the values of a particular sample's continuous variables and help clinicians implementing precision medicine with individualized risk factors of mortality.

3.1.3 Tree-based models capture non-linear relationships and important thresholds

Discovering non-linear relationships is important but challenging for epidemiological research using traditional linear models. J-shaped and U-shaped associations are two common and meaningful non-linear relationships [29]. However, as mentioned previously, linear models must use manually transformed features to capture non-linear relationships. Three common statistical approaches to modeling non-linear relationships using linear models are: polynomial/exponential modeling, categorical reparameterization, and splines [29]. For polynomial/exponential modeling and splines, non-linear relationships are identified and appropriately transformed. For categorical reparameterization, the continuous variable is divided into mutually exclusive categories, with each category represented by a binary variable. As an example, Suliman *et al.* used a linear model to capture the non-linear relationship between uric acid and mortality risk [30]. The authors divided

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

uric acid level into three categories and calculated the hazard ratio for each. Based on these categories, they showed a J-shaped mortality relationship for uric acid levels in patients with CKD stage 5 who were starting renal replacement therapy.

Unlike linear models, tree-based approaches can capture non-linear relationships without manually transforming or discretizing variables. For instance, we observe a U-shaped relationship between uric acid level and all-cause 5-year mortality predictions in Figure 3B. This relationship differs from the J-shaped one in from previous work, possibly because of categorization which loses important information about values within the categories. In summary, modeling non-linear relationships using linear models with categorical reparameterization has a few key disadvantages: (1) loss of information, (2) sensitivity to the choice of subgroups, and (3) loss of predictive power.

Additionally, discovering thresholds (i.e., inflection points beyond which changing a feature's value has diminishing returns) is significant in epidemiological analysis. However, it is hard to discover important thresholds using linear models, since linear models are not able to capture non-linear relationships. Figure 3C illustrates the main effect of urine albumin for the 5-year mortality model. The plot shows that 250 μg/mL is an important threshold: according to our model, increasing urine albumin generally increases 5-year mortality risk; however, urine albumin higher than this threshold has almost the same impact on mortality risk.

3.1.4 Tree-based models measure feature interaction effects

In contrast to linear models, tree-based models can naturally capture interaction effects by splitting on different features in the same tree. TreeExplainer extends local explanations to capture these interactions. In particular, SHAP dependence plots show how a feature's value (x-axis) impacts the prediction (y-axis) of each sample (Figure 3D); the color corresponds to the value of a second feature that may have an interaction effect with the feature we are plotting. Figure 3D highlights a specific interaction: the relationship of blood lead level to mortality presents differently for young and old individuals. Specifically, for those with blood lead higher than 0.1 μmol/L, younger individuals have a higher 5-year mortality risk than older individuals; the opposite is true for individuals with blood lead level lower than 0.1 μmol/L. This finding highlights how the non-linear interaction effects detected by our model open opportunities for further research.

SHAP dependence plots can be decomposed into main effects and interaction effects for each sample (Figure 3D-F). The plot of the SHAP interaction effect between age and gender (Figure 3G) shows a change in mortality risk between men and women for different age categories. From age 70 and older, men have a greater 5-year mortality risk than women. Although some previous work noted gender differences in mortality risk [31–33], it is difficult to illustrate the interaction effect of age and sex to mortality using traditional approaches. As another example, Figure 3H shows the SHAP interaction effects of cancer and gender with

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

blood lead level. For samples with blood lead levels higher than 0.11 μmol/L, individuals diagnosed with cancer have a higher 5-year mortality risk than individuals without cancer. Similarly, females have a higher 5-year mortality risk than males with blood lead levels higher than 0.24 μmol/L. The interaction effects of age and gender with blood lead level cannot be clearly captured without SHAP interaction values because being male or older generally increases mortality risk relative to being female or younger. In summary, the tree-based model captures interaction effects undetected by other approaches.

3.1.5 Tree-based models highlight different non-linear relationships between models

The relationship between each feature and mortality may change for models that predict different mortality outcomes or utilize different subsamples of the general population. For instance, comparing important features between a 1-year mortality prediction model and a 5-year mortality prediction model can reveal features that are only predictive of short-term mortality, not longer-term mortality (and vice versa). Likewise, we can compare models trained on distinct subpopulations, e.g., individuals younger than 40 years old and individuals older than 80. The differences between these models can help researchers identify risk factors relevant to each subpopulation. Comparing models in this way can provide epidemiological insights that may guide policy for specific at-risk populations.

Figure 3J–K shows that low serum sodium is protective for shorter term mortality (1 year) but harmful for longer term mortality (5 years) in our models. Hyponatremia is a low serum sodium concentration in the blood (<135 mmol/L). For the 1-year model, there is a positive relationship between serum sodium and mortality when serum sodium exceeds 135 mmol/L, which indicates higher serum sodium levels, especially higher than 139 mmol/L are associated with 1-year mortality. For 1-year mortality, only high serum sodium concentration increases mortality risk. Hyponatremia with negative SHAP values decreased mortality risk. However, the relationship differs completely in the 5-year mortality prediction model. Samples with low serum sodium have higher mortality risk, which means hyponatremia is associated with higher 5-year mortality risk. This type of insight, especially regarding the differences of non-linear trends, is not apparent using linear model.

3.2 Discoveries from 5-year mortality prediction

3.2.1 IMPACT identifies the magnitude, prevalence, and direction of each feature on mortality risk

Figure 4A shows a SHAP summary plot that displays the magnitude, prevalence, and direction of a feature's effect for many samples. We plot the top 20 most impactful features for the 5-year mortality prediction

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

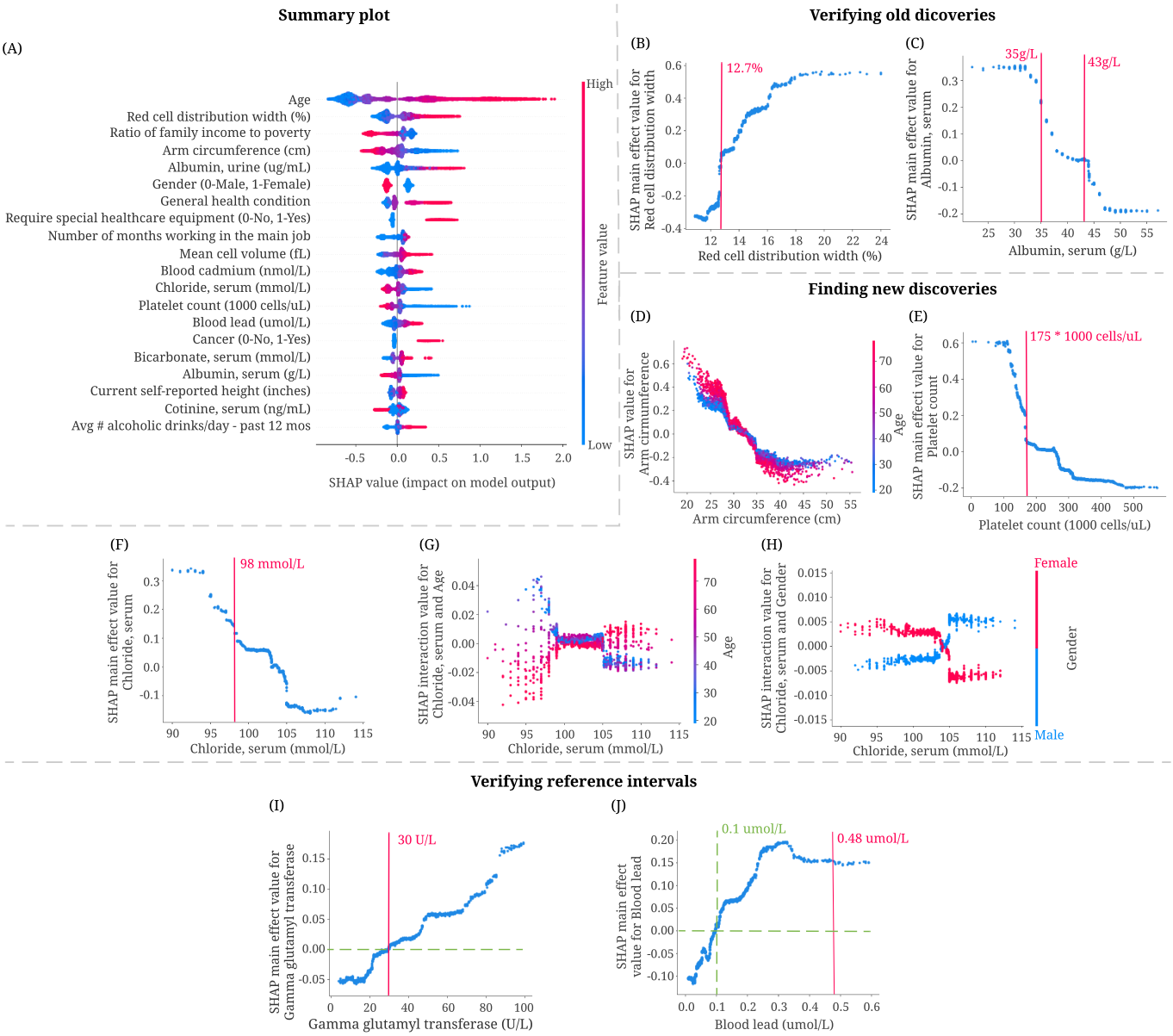


Figure 4: **Combining 5-year mortality prediction gradient boosted trees models and local explanations to achieve significant discoveries about the entire model and individual features**. (A) SHAP summary plot for the gradient boosted trees trained on the 5-year mortality prediction task (Section 3.2.1). The plot shows the most impactful features on prediction (ranked from most to least important) and the distribution of the impacts of each feature on the model output, which includes a set of plots where each dot corresponds to an individual. The colors represent feature values for numeric features: red for larger values, and blue for smaller. The thickness of the line that is comprised of individual dots is determined by the number of examples at a given value. A negative SHAP value (extending to the left) indicates reduced mortality risk, while a positive one (extending to the right) indicates increased mortality risk. (B,C) IMPACT can verify well-studied features associated with mortality (Section 3.2.2). (B) The main effect of red cell distribution width on 5-year mortality. (C) The main effect of serum albumin on 5-year mortality. (D-H) IMPACT can identify less well-studied features associated with mortality (Section 3.2.3). (D) SHAP value for arm circumference in 5-year mortality model. (E) The main effect of platelet count on 5-year mortality. (F) The main effect of serum chloride on 5-year mortality. (G) The SHAP interaction value of serum chloride vs. age in the 5-year mortality model. (H) The SHAP interaction value of serum chloride vs. gender in 5-year mortality model. (I-J) IMPACT can verify laboratory reference intervals (Section 3.2.4). (I) The main effect of gamma glutamyl transferase on 5-year mortality. (J) The main effect of blood lead on 5-year mortality.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

model. The long tails mean that the features are extremely predictive for some individuals but not others (e.g., urine albumin level). This summary plot provides an integrated explanation of the 5-year mortality model that indicates not only the overall ranking of the top 20 features but the specific effect of each feature on mortality risk.

3.2.2 IMPACT verifies well-studied features associated with mortality

Figure 4A shows the top 20 most important features for 5-year mortality prediction. Several features have previously been shown to be associated with mortality in epidemiological studies. Our results examine and support these studies' conclusions as well as surface additional discoveries, including novel features, thresholds, and non-linear relationships. As an example, red cell distribution width (RDW), which is the second most important feature of IMPACT's 5-year mortality model, has been shown to be strongly associated with mortality by many studies where higher RDW associated with increased mortality risks under several conditions [34–37]. In Figure 4B, we observe the positive relationship between RDW and risk of mortality; moreover, 12.7% is an important threshold over which RDW manifests a positive effect to mortality. Furthermore, serum albumin level's relation to mortality is also well-studied. It is the 17th most important feature in our 5-year model. Previous studies show that serum albumin is negatively associated with mortality risk [38–40]. The main effect shown in Figure 4C matches this trend. Furthermore, Corti *et al.* show that hypoalbuminemia (serum albumin level<35 g/L) was associated with a significantly increased risk of mortality compared to serum albumin levels greater than 43 g/L [38]. From the main effect plot in Figure 4C, we observe that 35 g/L and 43 g/L are indeed key inflection points: serum albumin levels lower that 43 g/L have a positive relationship with mortality prediction, while those around 35 g/L are associated with a dramatically increased mortality risk.

3.2.3 IMPACT identifies less well-studied features associated with mortality

Most of the top 20 most important features for 5-year mortality prediction models have been previously identified; however, some are less well-studied in existing epidemiological research. Three of these are arm circumference, platelet count and serum chloride level. The SHAP value of arm circumference in Figure 4D shows a negative relationship between arm circumference and 5-year mortality risk, especially for older people. IMPACT ranks arm circumference as the fourth most important feature for 5-year mortality prediction, with an importance ranking that significantly exceeds that of BMI (the 56th). This suggests that smaller arm circumference is more predictive than BMI for modeling mortality, as in [41]. More experiment results about this discovery can be found in Appendix A.

Next, Figure 4E shows the main effect of platelet count, the 13th most important feature for 5-year

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

mortality prediction. We see a negative relationship between platelet count and 5-year mortality. Moreover, $175 \times 1,000$ cells/μL is an important threshold; platelet count lower than that level is associated with dramatically increased mortality risk. Serum chloride is also inversely related to 5-year mortality, as shown in Figure 4F. The normal adult value for chloride is 98-106 mmol/L. From the figure, we observe serum chloride lower than 98 mmol/L can sharply increase mortality risk. In Figure 4G–H, we plot the interaction effect of age and sex with serum chloride level. The figures reveal that younger people and females with low serum chloride have higher mortality risk than older people and males. These trends differ from the main effect of age and sex: the interaction effect of age and serum chloride shows that early rather than late onset hypochloremia (low chloride) is more concerning to the model.

3.2.4 IMPACT verifies laboratory reference intervals

A laboratory test result is of little value and hard to interpret unless it is compared with the reference interval (RI) which is the range of values that is deemed normal for a physiologic measurement in healthy persons [42]. Reference intervals are the most common decision support tool to interpret patient laboratory test results and enable differentiation of healthy and unhealthy individuals [43, 44]. Hence, the quality of the RIs is as important as the quality of the result itself. RIs in use today are most commonly defined as the central 95% of laboratory test results in a reference population. Unfortunately, this definition doesn't take mortality or disease risk into consideration which may lead to misdiagnosis of laboratory results since reference intervals are used to differentiate unhealthy individuals. The SHAP values in IMPACT reflect the effects of different laboratory test results to mortality risk, which are able to verify the reference intervals.

From the main effect plot in Figure 4I, we observe that gamma glutamyl transferase (GGT) greater than 30 U/L has positive SHAP values that imply high values of GGT increases 5-year mortality risk. We also find that GGT levels in the range of 0 to 30 U/L are normal in adults[1]. IMPACT helps verify that the reference interval of GGT is appropriate and reasonable. The reference interval for blood lead is less than 0.48 µmol/L[2]. However, the main effect of blood lead in Figure 4J reveals that blood lead higher than 0.1 µmol/L increases 5-year mortality risk. The SHAP value of 0.48 µmol/L is very high and thus associated with increased mortality risk. Hence, IMPACT identifies that the upper bound of the blood lead reference interval might be too high.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

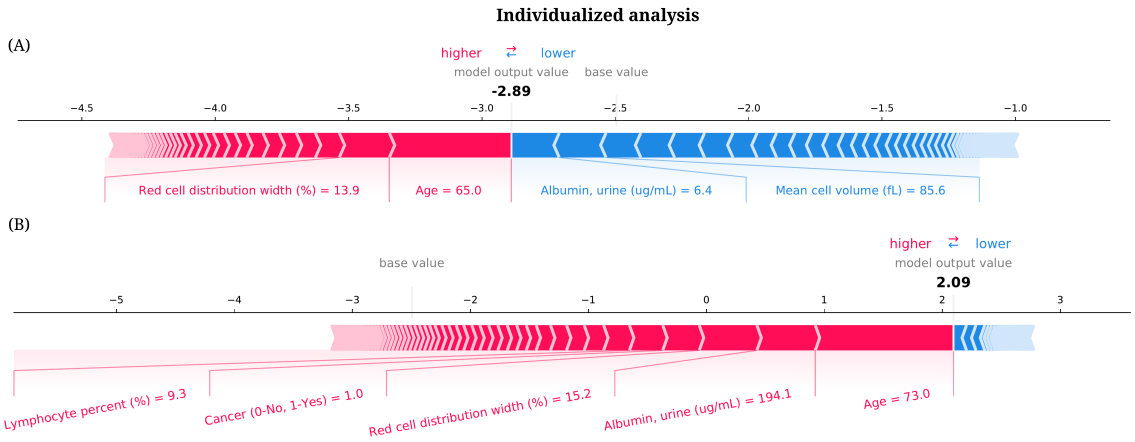


Figure 5: **Combining 5-year mortality prediction gradient boosted trees models and local explanations to achieve explanation for individual samples**. Tree-based mortality models can analyze individualized mortality predictions (Section 3.2.5). (A) The individualized explanation for an individual who is alive after 5 years. The output value is the prediction for that individual. The base value is the mean prediction (the value that would be predicted if we did not know any features for the current output). The features in red increase mortality risk, and those in blue decrease it. (B) The individualized explanation for a sample who is deceased after 5 years.

3.2.5 IMPACT analyzes individualized mortality predictions

TreeExplainer can help researchers analyze the prediction for each individual and illustrate each features' contribution to mortality risk. Figure 5 shows individualized explanations for two samples. The first sample (Figure 5A) was alive after 5 years. IMPACT predicted that the sample's 5-year mortality probability was 0.02. From the figure, we observe the mortality risk (i.e., the log-odds) is -2.89, lower than the average predicted risk (i.e., base value). There are features with positive attributions (i.e., the feature increases mortality risk prediction), such as low urine albumin concentration and normal mean cell volume, and negative attributions (i.e., the feature decreases mortality risk prediction), such as high age and high red cell distribution width. For this sample, the features that drive down mortality risk outweigh those that increase it. The second sample (Figure 5B) was deceased after 5 years, and the model's predicted mortality probability is 0.80. The individualized explanation shows a high mortality risk of 2.09, much higher than the average predicted risk. The top five features that increase this sample's risks are high age, high urine albumin concentration, red cell distribution width, diagnosis of cancer, and low lymphocyte percentage. For this sample, features that increase risk prediction dominate the others. These individualized explanations can help doctors give personalized treatment and implement precision medicines.

[1]https://www.webmd.com/hepatitis/ggt-test
[2]https://www.ucsfhealth.org/medical-tests/003360

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

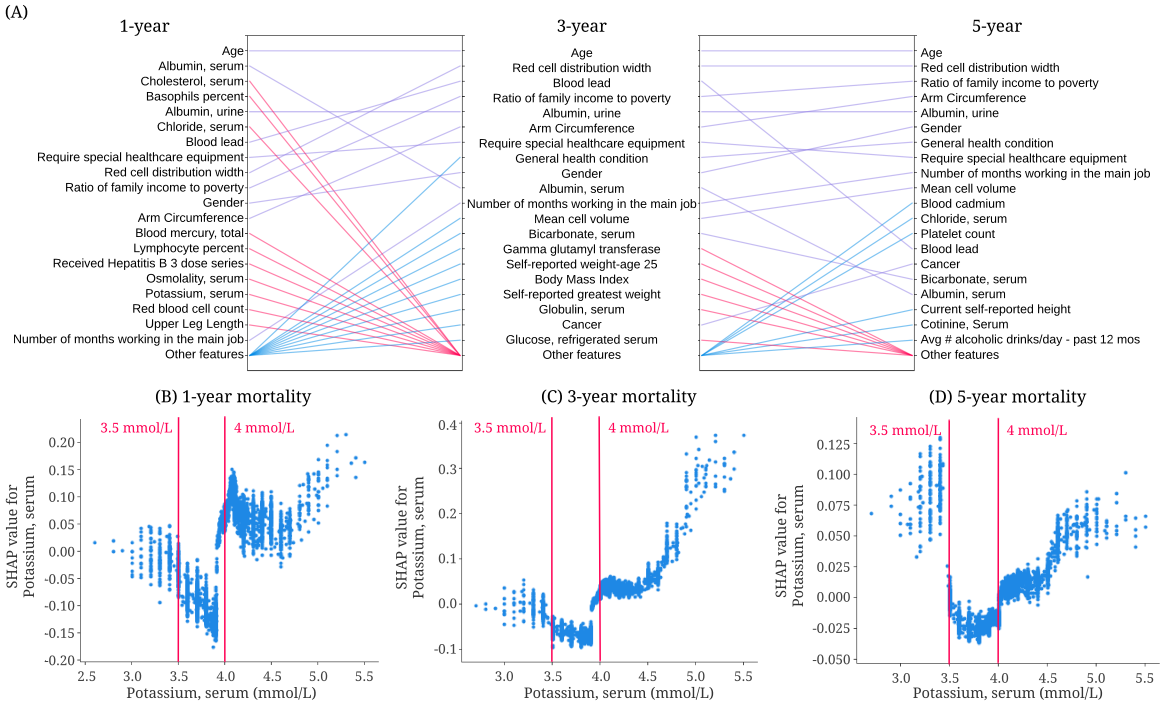


Figure 6: **Identifying important discoveries for mortality prediction from tree-based models using different follow-up times (Section 3.3.1).** (A) Relative importance of input features in 1-, 3-, and 5-year mortality models. For each model, the figure shows the 20 most important features on prediction (ordered by the importance). The purple line indicates that the feature is in the top 20 features of two models. Blue and red lines indicate the feature is in the top 20 features of one model, but not in the top 20 features of the other. (B) The SHAP value of serum potassium in the 1-year mortality model. (C) The SHAP value of serum potassium in the 3-year mortality model. (D) The SHAP value of serum potassium in the 5-year mortality model.

3.3 Discoveries for mortality prediction using different follow-up times and age groups

3.3.1 IMPACT identifies trends for 1-year, 3-year and 5-year mortality prediction models

Figure 6A shows the top 20 most important features and relative importance of input features in IMPACT's 1-year, 3-year and 5-year mortality prediction models. Feature importance rankings change significantly between three models. Some features are important for all three (e.g., age, gender, red cell distribution width and urine albumin level). For longer-term mortality prediction, some features become more important over time (e.g., platelet count whose importance ranking is 75 for the 1-year model and 13 for the 5-year model). Other features become less important over time (e.g., serum potassium whose importance ranking is 17 for the 1-year model and the 57 for 5-year model). This figure reveals the change of each feature's importance to shorter or longer term mortality prediction. These results can give the epidemiologists a more comprehensive understanding of all-cause mortality and help them investigate risk factors and make interventions further.

IMPACT across the 1, 3, 5-year mortality prediction models shows that high potassium is more harmful

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

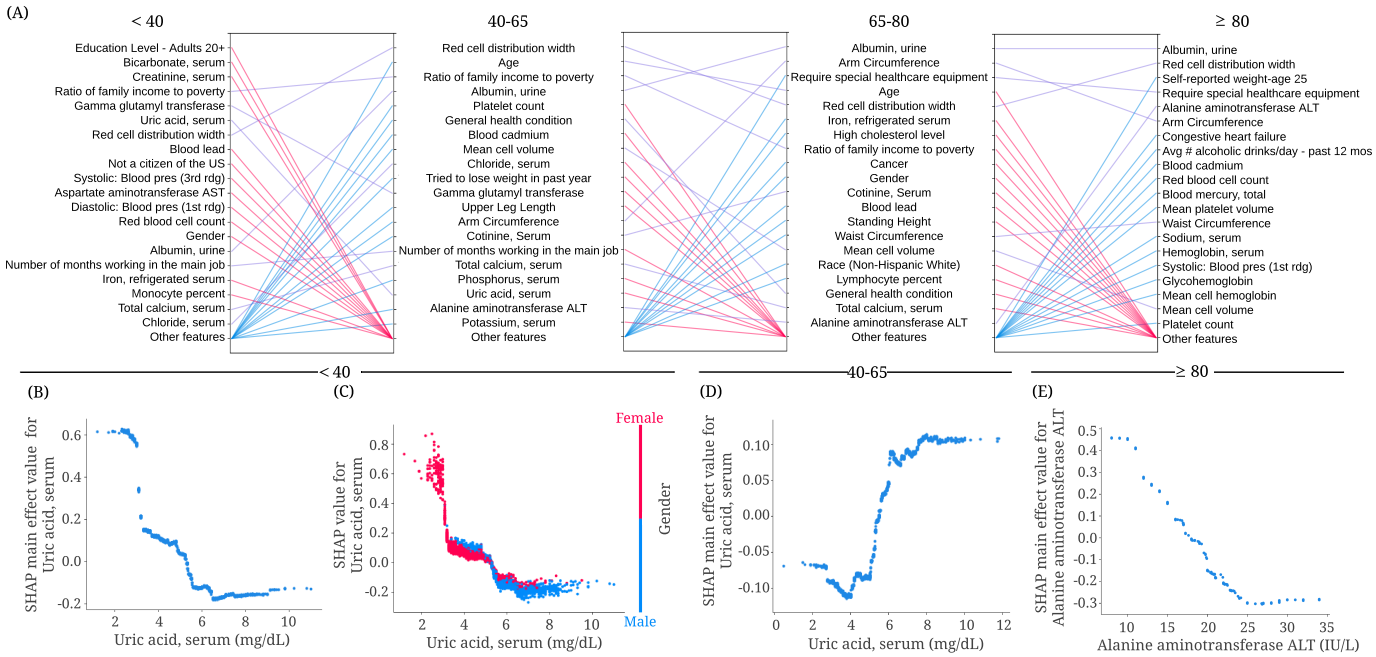


Figure 7: **Identifying important discoveries for mortality prediction in different age groups (Section 3.3.2).** (A) Relative importance of input features in <40, 40-65, 65-80 and ≥80 age groups. For each model, the figure shows the 20 most impactful features on prediction (ranked from most to least important). The purple line indicates that the feature is in the top 20 features of two models. Blue and red lines indicate the feature is in the top 20 features of one model, but not in the top 20 features of the other. (B) The main effect of serum uric acid on 5-year mortality in the <40 age group. (C) The SHAP value of serum uric acid in the <40 age group 5-year mortality model. (D) The main effect of serum uric acid on 5-year mortality in the 40-65 age group. (E) The main effect of alanine aminotransferase on 5-year mortality in the ≥80 age group.

for shorter term mortality where as low potassium is more harmful for longer term mortality. Figure 6B-D shows the SHAP value for serum potassium in IMPACT's 1-year, 3-year and 5-year mortality prediction models. We can see that serum potassium in the range of 3.5 mmol/L to 4.0 mmol/L has the lowest SHAP values which indicate lowest contribution to mortality risk. Serum potassium lower than 3.5 mmol/L and higher than 4.0 mmol/L are associated with increased mortality risk, a finding independently discussed in existing studies [45–47]. However, the trend differs across those models. For the 1-year and the 3-year model, hyperkalemia (high potassium) has a higher mortality risk than hypokalemia (low potassium). For the 5-year model, hypokalemia has the same or higher mortality risk than hyperkalemia.

3.3.2 IMPACT identifies important features for mortality prediction in different age groups

Figure 7A shows the top 20 most important features and relative importance of input features in 5-year mortality prediction models using different age groups (<40, 40-65, 65-80 and ≥80). We observe that the 20 most important features differ significantly across these models. Some features become more important for older subpopulations, such as alanine aminotransferase (ALT), the fifth most important feature in the

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

model using samples over 80 years. Figure 7E shows the main effect of ALT for age≥80, which shows the negative relationship between ALT and 5-year mortality. Some existing studies have found that alanine aminotransferase (ALT) is associated with all-cause mortality and some cause-specific mortality, such as cardiovascular disease, cancer, diabetes mellitus and liver disease [48, 49].

Moreover, some features are less important for older subpopulations than younger ones. One example is uric acid level, the sixth most important feature in the age<40 model and the 59th most important in age≥80 model. Figure 7B,C plot the main effect and SHAP value of uric acid in the age<40 model, showing that low uric acid levels increase mortality risk prediction. However, in the age 40-65 model, higher uric acid is associated with higher mortality risk (Figure 7D). Cho *et al.* concluded that both low and high uric acid levels were predictors of increased mortality [50]. Other work also shows that low uric acid in blood serum can injure the endothelium and induce oxidative stress-related disease, such as hypertension, diabetes mellitus and kidney disease [51, 52], and that hyperuricemia (high uric acid) is associated with various adverse health outcomes, including hypertension, metabolic syndrome, kidney disease, stroke, cardiovascular disease and cancer [53–56]. The numerous downstream effects of high uric acid and low uric acid might explain the different relationship between uric acid and mortality in different age groups. Moreover, the reference range of uric acid differs for males and females. The reference range is 2.4-6.0 mg/dL for females and 3.4-7.0 mg/dL for males. This difference is shown in Figure 7C, where women have lower uric acid, which can increase mortality risk.

4 Discussion

Here, we present IMPACT framework, an explainable tree-based model for mortality prediction using a wide array of demographic, laboratory, examination and questionnaire variables from 47,261 NHANES (1999-2014) participants. To interpret the "black box" model, we use TreeExplainer to calculate SHAP values. To our knowledge, IMPACT is the first study that combines complex ML models and state-of-the-art local explanation methods to predict mortality in epidemiological studies. We illustrate the advantages of tree-based complex ML models for epidemiological studies.

Using the most overlapping features in NHANES 1999-2014, we build tree-based mortality prediction models and explore the effect of those features on mortality for different follow-up times and age groups. Importantly, we demonstrate the value and significance of explaining complex ML prognostic models. In epidemiological studies, high prediction performance is not enough; instead, interpretation of the model is essential for drawing epidemiological conclusions [4, 11]. IMPACT combines high-accuracy complex ML model with interpretable explanations, which allows us to capture the effect of variables on mortality and include

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

non-linear effects as well as interactions between variables that are difficult to identify using linear models. Local explanations from TreeExplainer show the contribution of each feature as well as their interactions. These explanations allow us to verify well-studied findings as well as identify new ones.

Limitations: First, the relationships and interactions detected by our model cannot be claimed to be causal. The purpose of this study is not to address causality, but rather to understand complex interrelations between mortality and individual's features. Further analysis using causal epidemiological methods must be undertaken before claiming causality. Second, our study is performed on NHANES 1999-2014, which is designed to assess the health status of participants in the United States. The conclusions and discoveries can generalize to other populations only when the distribution of variables and mortality rates are similar to those in the U.S. as a whole. Further external validation of our mortality models on datasets from non-U.S. populations should be undertaken to increase the generalizability of these findings.

Future implications: To improve our model, a mortality prediction tool that calculates a mortality risk score and provides individual explanations can be developed. Such a tool could help individuals improve self-awareness of their health status and help clinicians identify patients with high mortality risk to target with specific interventions.

Prognosis research using complex ML models will likely increase over the coming years as ML techniques continue to rapidly develop. However, "black box" ML models that predict without explaining, are difficult for clinicians to trust and use to extract important information. Therefore, the combination of complex ML models and 'explainable artificial intelligence' (XAI) is necessary and urgent. IMPACT takes a significant step towards XAI for mortality prediction. The improvement in predictive accuracy and explanation of complex ML models shown in this study should be further explored for predicting other epidemiological outcomes.

5 Conclusion

This study demonstrates the advantages of interpretable complex machine learning for all-cause mortality prediction within an epidemiological study design. It illustrates how the combination of complex machine learning model and interpretable methods can improve prognostic research, verify well-studied findings and explore new discoveries.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* **316,** 2402–2410 (2016).

2. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* **24,** 1559–1567 (2018).

3. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118 (2017).

4. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one* **12,** e0174944 (2017).

5. Wang, H. *et al.* Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM transactions on computational biology and bioinformatics* **15,** 1968–1978 (2018).

6. Chau, J. Y. *et al.* Daily sitting time and all-cause mortality: a meta-analysis. *PloS one* **8,** e80000 (2013).

7. Kuk, J. L. *et al.* Visceral fat is an independent predictor of all-cause mortality in men. *Obesity* **14,** 336–341 (2006).

8. Marshall, N. S. *et al.* Sleep apnea as an independent risk factor for all-cause mortality: the Busselton Health Study. *Sleep* **31,** 1079–1085 (2008).

9. Walter, S. *et al.* Genetic, physiological, and lifestyle predictors of mortality in the general population. *American journal of public health* **102,** e3–e10 (2012).

10. Ganna, A. & Ingelsson, E. 5 year mortality predictors in 498 103 UK Biobank participants: A prospective population-based study. *The Lancet* **386,** 533–540. ISSN: 1474547X. `http://dx.doi.org/10.1016/S0140-6736(15)60175-1` (2015).

11. Weng, S. F., Vaz, L., Qureshi, N. & Kai, J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PloS one* **14,** e0214365 (2019).

12. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2,** 2522–5839 (2020).

13. Hilton, C. B. *et al.* Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *NPJ Digital Medicine* **3,** 1–8 (2020).

14. Vaid, A. *et al.* Machine Learning to Predict Mortality and Critical Events in COVID-19 Positive New York City Patients. *medRxiv* (2020).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

15. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118 (2012).

16. Krauss, C., Do, X. A. & Huck, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research* **259,** 689–702 (2017).

17. Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E. & Petrovski, K. R. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models. *Computers in biology and medicine* **114,** 103456 (2019).

18. Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system* in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), 785–794.

19. Lundberg, S. M. & Lee, S.-I. *A unified approach to interpreting model predictions* in *Advances in neural information processing systems* (2017), 4765–4774.

20. Molnar, C. *Interpretable machine learning* ISBN: 0244768528 (Lulu. com, 2019).

21. Yang, Y., Morillo, I. G. & Hospedales, T. M. Deep neural decision trees. *arXiv preprint arXiv:1806.06988* (2018).

22. Fan, C., Liu, D., Huang, R., Chen, Z. & Deng, L. *PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility* in *Bmc Bioinformatics* **17** (2016), 85–95.

23. Torlay, L., Perrone-Bertolotti, M., Thomas, E. & Baciu, M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain informatics* **4,** 159–169 (2017).

24. Ren, X., Guo, H., Li, S., Wang, S. & Li, J. *A novel image classification method with CNN-XGBoost model* in *International Workshop on Digital Watermarking* (2017), 378–390.

25. Zhang, D. *et al.* A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access* **6,** 21020–21031 (2018).

26. Lustberg, M. & Silbergeld, E. Blood lead levels and mortality. *Archives of internal medicine* **162,** 2443–2449. ISSN: 0003-9926 (2002).

27. Menke, A., Muntner, P., Batuman, V., Silbergeld, E. K. & Guallar, E. Blood lead below 0.48 mmol/L (10 mg/dL) and mortality among US adults. *Circulation* **114,** 1388–1394 (2006).

28. Schober, S. E., Mirel, L. B., Graubard, B. I., Brody, D. J. & Flegal, K. M. Blood lead levels and death from all causes, cardiovascular disease, and cancer: results from the NHANES III mortality study. *Environmental health perspectives* **114,** 1538–1541. ISSN: 0091-6765 (2006).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

29. May, S. & Bigelow, C. Modeling nonlinear dose-response relationships in epidemiologic studies: statistical approaches and practical challenges. *Dose-Response* **3,** dose–response. ISSN: 1559-3258 (2005).

30. Suliman, M. E. *et al.* J-shaped mortality relationship for uric acid in CKD. *American Journal of Kidney Diseases* **48,** 761–771. ISSN: 0272-6386 (2006).

31. Case, A. & Paxson, C. Sex differences in morbidity and mortality. *Demography* **42,** 189–214. ISSN: 0070-3370 (2005).

32. Oksuzyan, A., Juel, K., Vaupel, J. W. & Christensen, K. Men: good health and high mortality. Sex differences in health and aging. *Aging clinical and experimental research* **20,** 91–102. ISSN: 1594-0667 (2008).

33. Wingard, D. L. The sex differential in mortality rates: demographic and behavioral factors. *American Journal of Epidemiology* **115,** 205–216. ISSN: 1476-6256 (1982).

34. Felker, G. M. *et al.* Red cell distribution width as a novel prognostic marker in heart failure: data from the CHARM Program and the Duke Databank. *Journal of the American College of Cardiology* **50,** 40–47. ISSN: 0735-1097 (2007).

35. Patel, K. V., Ferrucci, L., Ershler, W. B., Longo, D. L. & Guralnik, J. M. Red blood cell distribution width and the risk of death in middle-aged and older adults. *Archives of internal medicine* **169,** 515–523. ISSN: 0003-9926 (2009).

36. Patel, K. V. *et al.* Red cell distribution width and mortality in older adults: a meta-analysis. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* **65,** 258–265. ISSN: 1758-535X (2010).

37. Perlstein, T. S., Weuve, J., Pfeffer, M. A. & Beckman, J. A. Red blood cell distribution width and mortality risk in a community-based prospective cohort. *Archives of internal medicine* **169,** 588–594. ISSN: 0003-9926 (2009).

38. Corti, M.-C., Guralnik, J. M., Salive, M. E. & Sorkin, J. D. Serum albumin level and physical disability as predictors of mortality in older persons. *Jama* **272,** 1036–1042. ISSN: 0098-7484 (1994).

39. Goldwasser, P. & Feldman, J. Association of serum albumin and mortality risk. *Journal of clinical epidemiology* **50,** 693–703. ISSN: 0895-4356 (1997).

40. Phillips, A., Shaper, A. G. & Whincup, P. Association between serum albumin and mortality from cardiovascular disease, cancer, and other causes. *The Lancet* **334,** 1434–1436. ISSN: 0140-6736 (1989).

41. Tsai, A. C. & Chang, T.-L. The effectiveness of BMI, calf circumference and mid-arm circumference in predicting subsequent mortality risk in elderly Taiwanese. *British Journal of Nutrition* **105,** 275–281. ISSN: 1475-2662 (2011).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

42. Katayev, A., Balciza, C. & Seccombe, D. W. Establishing reference intervals for clinical laboratory test results: is there a better way? *American journal of clinical pathology* **133,** 180–186 (2010).

43. Ozarda, Y., Higgins, V. & Adeli, K. Verification of reference intervals in routine clinical laboratories: practical challenges and recommendations. *Clinical Chemistry and Laboratory Medicine (CCLM)* **57,** 30–37 (2018).

44. Jones, G. & Barker, A. Reference intervals. *The Clinical Biochemist Reviews* **29,** S93 (2008).

45. Ahmed, A. *et al.* A propensity-matched study of the association of low serum potassium levels and mortality in chronic heart failure. *European heart journal* **28,** 1334–1343. ISSN: 1522-9645 (2007).

46. Goyal, A. *et al.* Serum potassium levels and mortality in acute myocardial infarction. *Jama* **307,** 157–164. ISSN: 0098-7484 (2012).

47. Nakhoul, G. N. *et al.* Serum potassium, end-stage renal disease and mortality in chronic kidney disease. *American journal of nephrology* **41,** 456–463. ISSN: 0250-8095 (2015).

48. Ruhl, C. E. & Everhart, J. E. The association of low serum alanine aminotransferase activity with mortality in the US population. *American journal of epidemiology* **178,** 1702–1711. ISSN: 1476-6256 (2013).

49. Vespasiani-Gentilucci, U. *et al.* Low alanine aminotransferase levels in the elderly population: frailty, disability, sarcopenia, and reduced survival. *The Journals of Gerontology: Series A* **73,** 925–930. ISSN: 1079-5006 (2018).

50. Cho, S. K., Chang, Y., Kim, I. & Ryu, S. U-Shaped Association Between Serum Uric Acid Level and Risk of Mortality: A Cohort Study. *Arthritis & Rheumatology* **70,** 1122–1132. ISSN: 2326-5191 (2018).

51. Stiburkova, B., Stekrova, J., Nakamura, M. & Ichida, K. Hereditary renal Hypouricemia type 1 and autosomal dominant polycystic kidney disease. *The American journal of the medical sciences* **350,** 268–271. ISSN: 0002-9629 (2015).

52. Wakasugi, M. *et al.* Association between hypouricemia and reduced kidney function: a cross-sectional population-based study in Japan. *American journal of nephrology* **41,** 138–146. ISSN: 0250-8095 (2015).

53. Feig, D. I., Kang, D.-H. & Johnson, R. J. Uric acid and cardiovascular risk. *New England Journal of Medicine* **359,** 1811–1821. ISSN: 0028-4793 (2008).

54. Ford, E. S., Li, C., Cook, S. & Choi, H. K. Serum concentrations of uric acid and the metabolic syndrome among US children and adolescents. *Circulation* **115,** 2526–2532. ISSN: 0009-7322 (2007).

55. Lehto, S., Niskanen, L., Ronnemaa, T. & Laakso, M. Serum uric acid is a strong predictor of stroke in patients with non–insulin-dependent diabetes mellitus. *Stroke* **29,** 635–639. ISSN: 0039-2499 (1998).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

56. Strasak, A. M. *et al.* Serum uric acid and risk of cancer mortality in a large prospective male cohort. *Cancer causes & control* **18,** 1021–1029. ISSN: 0957-5243 (2007).

57. Coelho, A. K., Rocha, F. L. & Fausto, M. A. Prevalence of undernutrition in elderly patients hospitalized in a geriatric unit in Belo Horizonte, MG, Brazil. *Nutrition* **22,** 1005–1011 (2006).

58. Kim, Y. H. & So, W.-Y. A low arm and leg muscle mass to total body weight ratio is associated with an increased prevalence of metabolic syndrome: The Korea National Health and Nutrition Examination Survey 2010-2011. *Technology and Health Care* **24,** 655–663 (2016).

59. Wijnhoven, H. A. *et al.* Low mid-upper arm circumference, calf circumference, and body mass index and mortality in older persons. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* **65,** 1107–1114 (2010).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A More analyses of BMI-related features

Feature	Correlation with BMI
Waist circumference	0.8915
Arm circumference	0.8815
Weight	0.8748
Current self-reported weight	0.8256
Self-reported weight - 1 year ago	0.7673
Self-reported greatest weight	0.7559

Figure 8: The correlations of six BMI-related features and BMI.

In our experiment results, we found that IMPACT ranks arm circumference as the fourth most important feature for 5-year mortality prediction, with an importance ranking that significantly exceeds that of BMI (the 56th). To further analyze this discovery, we do more experiments on BMI-related features. Besides BMI, we select 6 more features that are highly correlated with BMI, including waist circumference, arm circumference, weight, current self-reported weight, self-reported weight 1 year ago and self-reported greatest weight. The correlations between BMI and these features are shown in Figure 8.

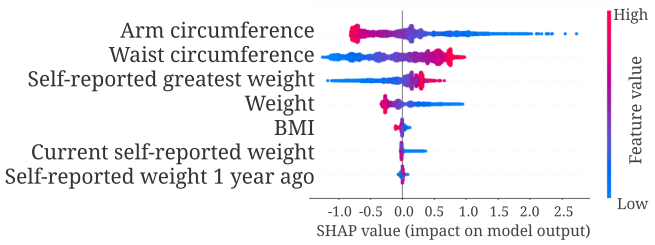


Figure 9: SHAP summary plot of the 5-year mortality prediction GBT model trained on BMI-related features.

Feature	AUROC
Arm circumference	0.6010
Weight	0.5476
BMI	0.5369
Current self-reported weight	0.5344
Waist circumference	0.5331
Self-reported greatest weight	0.5288
Self-reported weight - 1 year ago	0.5246

Figure 10: The AUROCs of uni-variate 5-year mortality prediction GBT models trained on each BMI-related featurs.

Firstly, we trained a multi-variate GBT model to predict 5-year mortality using these BMI-related features. From the summary plot in Figure 9, we can see that arm circumference is the most important feature for 5-year mortality prediction in all BMI-related features. What's more, waist circumference, self-reported greatest weight and weight are more important than BMI. To further investigate the predictive power of

medRxiv preprint doi: https://doi.org/10.1101/2021.01.20.21250135; this version posted January 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

BMI-related features on mortality, we train a uni-variate GBT model for each feature. The AUROCs of these models are shown in Figure 10. The model that uses arm circumference achieves the highest AUROC which is much higher than the AUROC of model that used BMI. These experiment results demonstrate that the arm circumference is more predictive than BMI for 5-year mortality prediction.

Arm circumference has been found to be useful for evaluating health and nutritional status of elder people [57]. Previous papers also mentioned that arm circumference is associated with metabolic syndrome [58] and mortality [59]. Our results also indicate that the arm circumference is more predictive than BMI for mortality prediction. As arm circumference is simpler to measure than BMI and for people who have difficulty in standing erectly, arm circumference is accessible, we can use arm circumference instead of BMI in further mortality prediction models.