Interpretations of Predictive Models for Lifestyle-related Diseases at Multiple Time Intervals

Yuki Oba[1] ✉, Taro Tezuka[2][0000−0002−5628−9961], Masaru Sanuki[3][0000−0003−4247−539X], and Yukiko Wagatsuma[3][0000−0002−8056−9998]

[1] Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan s2230178@s.tsukuba.ac.jp
[2] Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Japan tezuka@iit.tsukuba.ac.jp
[3] Faculty of Medicine, University of Tsukuba, Tsukuba, Japan {sanuki,ywagats}@md.tsukuba.ac.jp

Abstract. Health screening is practiced in many countries to find asymptotic patients of diseases. There is a possibility that applying machine learning to health screening datasets enables predicting future medical conditions. We extend this approach by introducing interpretable machine learning and determining health screening items (attributes) that contribute to detecting lifestyle-related diseases in their early stages. Furthermore, we determine how contributing attributes change within one to four years of time. We target diabetes and chronic kidney disease (CKD), which are among the most common lifestyle-related diseases. We trained predictive models using XGBoost and estimated each attribute's contribution levels using SHapley Additive exPlanations (SHAP). The results indicated that numerous attributes drastically change their levels of contribution over time. Many of the results matched our medical knowledge, but we also obtained unexpected outcomes. For example, we found that for predicting HbA1c and creatinine, which are indicators of diabetes and CKD, respectively, the contribution from alanine transaminase goes up as the time interval lengthens. Such findings can provide insights into the underlying mechanisms of how lifestyle-related diseases aggravate.

Keywords: Interpretable machine learning · data-driven medicine · health screening · disease prediction · tabular data

1 Introduction

Health screening is practiced in many countries to find asymptotic patients of diseases. It has become increasingly important to detect early symptoms of lifestyle-related diseases through health screening due to their increasing rate of patients among diseases. Machine learning provides a promising approach for making such predictions. In addition to finding asymptotic patients, there is high potential in health screening data. Medical researchers can obtains insights from

analyzing health screening data. Such an approach is often called data-driven medicine.

Data-driven medicine has been gaining much popularity due to its potentially high impact on medical practices and drug discovery. In the field of artificial intelligence (AI) research, interpretable machine learning provides a good measure for data-driven medicine. Medical researchers can gain an understanding from the insights provided by the interpretations of machine learning models into the underlying mechanisms of diseases. Clinically, an interpretation can suggest testing other examination items not included in the original health screening records to further understand the patient's condition.

In this paper, we introduce the aspect of time into interpretable machine learning to health screening. Existing works on predictions and interpretations have conducted analyses only at a single time interval, that is, using a specific time interval between features and the target attribute. However, as a lifestyle-related disease develops over time, different test items are affected. The change can be observed by comparing attributes contributing to making predictions at different points in time. A number of attributes may contribute to making long-term predictions, while others are useful for short-term predictions. Such a difference may have been known to clinicians, but it was difficult to observe it quantitatively. Interpretable machine learning on health screening records can now provide a suitable means for such an analysis.

To see the dynamics of contributing attributes, we observed the differences in attributes that help make predictions at time intervals between one and four years. Specifically, we trained models using XGBoost and ranked the attributes in accordance with their contributions using SHapley Additive exPlanations (SHAP) [11].

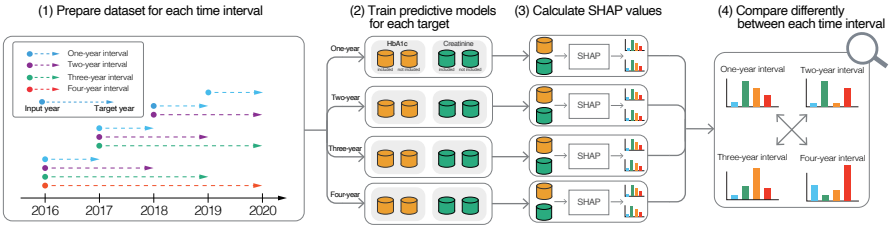


Fig. 1. The framework for interpreting predictive models presented in this paper.

We chose type 2 diabetes and chronic kidney disease (CKD) as the target diseases because they are the most common among lifestyle-related diseases. From the experiments, we found that the contributing attributes significantly differ depending on time intervals. Many attributes that contributed to predicting the disease indicators one year later were different from those four years later. Our results show that the interpretation of machine learning results can

effectively uncover such dynamics and provide more insights into how lifestyle-related diseases aggravate. This can also help patients and clinicians by enabling them to diagnose diseases at early stages. In addition, our approach provides a framework for applying interpretable machine learning to data-driven medicine in general and can contribute to the development of evidence-based medicine for lifestyle-related diseases. Figure 1 shows the overview of our experiments. The main contributions of the paper are as follows.

- We created a high-performance predictive model of lifestyle-related diseases for asymptomatic healthy patients based on a large-scale health screening dataset.
- We implemented a SHAP-based interpretation system for finding attributes that contribute to making predictions for the aggravation of diabetes and CKD.
- We found that different attributes contribute to the aggravation of diabetes and CKD depending on the time span being considered.

Our code is available at `https://github.com/itumizu/interpretation_at_multiple_time_intervals`.

2 Related work

2.1 Prediction of diabetes stages using medical records

Many studies proposed models to predict the onset of diabetes using medical records. Model building methods range from classical statistical approaches to modern machine learning-based techniques. For predictive models and methodologies, Kavakiotis *et al.* is a good survey on machine learning and data mining for diabetes research.

For classical statistical approaches, Sisodia *et al.* trained models such as decision trees and support vector machines (SVMs) to predict whether a patient would develop diabetes [19]. Choi *et al.* proposed a predictive model for type 2 diabetes using electronic medical records [4]. They used logistic regression, linear discriminant analysis, quadratic discriminant analysis, and k-nearest neighbor. Furthermore, Dagliati *et al.* applied a random forest and logistic regression on electronic health records to predict the onset of diabetes complications [5]. As modern machine learning techniques, Lai *et al.* used logistic regression and a gradient boosting machine to predict type 2 diabetes [10]. Most studies only focused on making predictions. However, several studies analyzed the implications of trained predictive models. For example, Manini *et al.* used a Bayesian network to investigate a causal structure for clinical complications in type 1 diabetes [13].

In most existing work, models were trained by medical data obtained from electronic medical records in hospitals. In contrast, our dataset comes from annual medical checkups. Additionally, since our dataset contains data from healthy subjects, it has an advantage in investigating the early stages of diabetes.

2.2 Prediction of chronic kidney diseases stages using medical records

CKD is known to cause and be caused by other diseases. Many studies aim to determine whether hospitalized patients having CKD later develop other diseases as a complication. Tangri *et al.* predicted whether patients with CKD would deteriorate to renal failure [21]. They used the Cox proportional hazards model to analyze the factors associated with CKD progression. Kunwar *et al.* built a model to classify whether a patient has CKD using a naive Bayesian method and a neural network and compared the performance [9]. Wang *et al.* used health checkup data to predict the risk of CKD using random forest, XGBoost, and ResNet by using the regression of creatinine levels [23]. Moreno-Sanchez proposed a model for early detection of CKD by combining AdaBoost and decision trees [15]. They analyzed the relationship between the feature importance obtained from the constructed model and the attributes entered into the model. For dealing with CKD complications, Ravizza *et al.* used logistic regression with electronic medical record data to predict the risk of developing CKD in patients with diabetes [17]. Belur Nagaraj *et al.* proposed a model to identify patients with type 2 diabetes who will develop end-stage renal failure in the future [2].

CKD progresses slowly due to lifestyle-related effects. Therefore, there may be no subjective symptoms until the disease becomes severe. Prediction of the long-term progression of the disease and detection of signs will help prevent the onset of the disease. However, existing studies almost entirely aimed at predicting the onset of CKD for patients already having certain diseases rather than targeting healthy subjects.

2.3 Interpretable prediction of diseases

Some studies have introduced interpretability into their prediction. Xie *et al.* identified two new risk factors for type 2 diabetes by training predictive models [24]. Their analysis included an SVM, random forest, and neural network. Using time-series data, Park *et al.* proposed to use deep attention networks to make interpretable predictions for vascular diseases [16]. Their model was based on an RNN, which is appropriate for long-time sequence data, but it may not suit the health screening records that we target. For interpretation at different time intervals, Shakeri *et al.* compared attributes that contribute to the prediction of sepsis onset at two and six hours using SHAP [18].

Broome *et al.* reviewed the status of machine learning and AI in making decisions regarding diabetes care. They covered numerous use cases of machine learning in identifying pre-diabetes patients, automated insulin dosing systems, and customized meal and lifestyle recommendations. Finally, Dankwa-Mullan *et al.* surveyed various ways AI can be used for diabetes care, discussing how interpretable models are critical for many applications [6].

Existing studies have added interpretability to conventional methods. On the other hand, few studies have attempted to use interpretability to find new relationships. Interpretable predictions lead to understanding the model's behavior

and why the predicted results were obtained. Moreover, interpretation of machine learning models can find abstract connections between the data learned by the model and the prediction target. Our method uses modern machine learning and interpretation techniques to analyze the relationship between the data and the predictor, focusing on the difference in the time intervals.

3 Dataset

3.1 Structure and attributes

We used a medical checkup dataset collected by a regional health care center in Mito Kyodo General Hospital in Japan. The dataset consists of three annual medical checkup records. The number of samples (participants) for 2016, 2017, 2018, 2019, and 2020 was 4,133, 4,261, 4,270, 4,015, and 4,367, respectively.

We conducted a prediction task with time intervals ranging from one to four years. We used records from participants who took medical checkups in both input and target years.

The number of participants in each combination were 2,396 for 2016–2017, 2,527 for 2017–2018, 2,511 for 2018-2019, 2,701 for 2019-2020, 2,140 for 2016-2018, 2,179 for 2017-2019, 2,531 for 2018-2020, 1,896 for 2016-2019, 2,274 for 2017-2020 and 1,975 for 2016-2020. We removed attributes that were missing in over 95% of the participants.

We trained our model using the 38 remaining attributes: age, sex, height, weight, waist circumference, body mass index (BMI), systolic blood pressure, diastolic blood pressure, total cholesterol, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, fasting blood sugar (FBS), hemoglobin A1c (HbA1c), status of diabetes mellitus, hemoglobin, red blood cell count, hematocrit, white blood cell count, uric acid, hematuria (blood in urine), urine protein, urine sugar, fecal occult blood for day 1, fecal occult blood for day 2, neutral fat, cholinesterase, creatinine, albumin, alanine transaminase, aspartate transaminase, γ-glutamyl transpeptidase, C-reactive protein, electrocardiogram, abdominal echo, chest X-ray, status of gastric intestinal series, ophthalmology, and serum abnormalities. In addition, there are answers to the 20 self-administered questions [14] shown in Table 1. We did not include Q13 and Q16 in the attributes we used because the questions have changed since 2018, and responses to the same questions are no longer available.

Sex was selected from male or female. Electrocardiogram, abdominal echo, chest X-ray, status of gastric intestinal series, ophthalmology, serum abnormalities were selected from "nothing particular," "mild abnormality, "follow-up," "requires treatment," "requires further testing," and "under medical treatment." Fecal occult blood for day 1 and day 2 were selected from "negative," "positive," and "missing." Urine protein and urine sugar were selected from (-), (+-), (+), (2+), (3+), (4+), (5+). Q1–Q12, Q14–Q15, Q17, Q20, and Q22 were yes/no questions. Q18 was selected from "every day," "sometimes," and "none." Q19 was selected from "less than 180 ml," "180–360 ml," "360–540 ml," and "more than

Table 1. Self-administered questions in our medical checkup dataset

Number	Question
Q1	Using anti-hypertensive drug
Q2	Using insulin injection or antidiabetic (hypoglycemic) drug
Q3	Using anti-cholesteremic agent
Q4	Have you ever been diagnosed as having a stroke (cerebral hemorrhage or infarction) by a physician or had medical treatment?
Q5	Have you ever been diagnosed as having heart disease (angina pectoris or myocardial infarction) by a physician or had medical treatment?
Q6	Have you ever been diagnosed as having chronic renal failure by a physician or got artificial dialysis?
Q7	Have you ever been diagnosed as having anemia?
Q8	Have you smoked in the last month?
Q9	Have you put on weight by 10 kg since your 20s?
Q10	Have you exercised for more than 30 minutes each time, for more than two times per week, and for more than one year?
Q11	Do you walk daily or do other physical activity equal to walking for more than 1 hour per day?
Q12	Do you walk faster than those in the same age group as you?
Q14	Do you eat faster than others?
Q15	Do you have dinner within two hours before going to bed more than three times a week?
Q17	Do you skip breakfast more than three times a week?
Q18	How often do you drink alcohol (such as sake, shochu, beer, whisky, etc.)?
Q19	When drinking, how much alcohol do you consume?
Q20	Do you sleep enough?
Q21	Do you want to improve your life style (life habit) such as exercise or eating?
Q22	If you have any chance to get health guidance on improving your life style (life habit), will you use it?

540 ml," where arbitrary alcoholic drinks was quantified by converting to sake containing the same amount of alcohol. Q21 was selected from "I am not planning on improving," "I would like to try," "I am starting," "I am improving (less than six months)", and "I am improving (more than six months)."

3.2 Ethical considerations

Annual medical examinations are conducted along with the Japanese Industry Safety and Health Act and are performed to facilitate lifestyle change and early disease diagnosis, which in turn would lower health expenditure and improve quality of life. This study was reviewed and approved by the ethics review committee of the authors' institution and conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from each participant.

4 Method

4.1 Target attributes

We used HbA1c and creatinine as target attributes because they are commonly-used indicators of diabetes and CKD, respectively. We refer to attributes used for predicting as "features." They are the inputs to predictive models. In statistical terms, features are independent variables, and target attributes are dependent variables.

HbA1c is widely used as a criterion for conducting a diabetes diagnosis. Creatinine is a metabolite created by energy production in muscles, and high amounts of it are found in patients with CKD. The estimated glomerular filtration rate (eGFR) is also widely used for the diagnosis of CKD. Because the eGFR can be computed deterministically from creatinine, age, sex, and race, we aimed at predicting the amount of creatinine.

4.2 Prediction tasks

The models were trained using features from a single year. Because our dataset contains health screening records from 2016 to 2020, we selected two years from 2016 to 2020 and used the latter half of the years chosen as the prediction target. For example, in one experiment, features from 2016 were used to predict the target attribute in 2020. In another experiment, features from 2017 were used to predict the target attribute in 2020. For each pair, features from the earlier year were used to predict the target attribute in the latter year.

To find attributes that contribute in making predictions in different ways, we conducted two types of experiments: (1) Training with the target attribute, together with strongly relevant attributes, from earlier years included as features, and (2) training without the target attribute and strongly relevant attributes from earlier years removed from features. In the latter type of experiments, HbA1c, FBS, and status of diabetes mellitus were removed from the attributes when predicting HbA1c. For predicting creatinine, only creatinine from earlier years was removed from features.

4.3 Preprocessing

In our dataset, each participant is labeled with one of six possible stages of diabetes, namely *nothing particular*, *mild abnormality*, *follow-up*, *requires treatment*, *requires further testing*, and *under medical treatment*. These stages were defined by the Japan Society of Ningen Dock[4]. For predicting HbA1c, we only used data from participants whose stage in the input year is in *nothing particular*, *mild abnormality*, or *follow-up*.

When predicting creatinine, we used the eGFR to filter out a number of participants. To measure the kidneys' filtering capacity, the eGFR, which is

[4] https://www.ningen-dock.jp/wp/wp-content/uploads/2018/06/Criteria-category.pdf

calculated from creatinine and age, is commonly used. We calculated the eGFR using the following formula 1 defined by the Japanese Society of Nephrology [8].

$$eGFR = 194 \times Creatinine^{-1.094} \times Age^{-0.287} (\times 0.739 \text{ if female}) \quad (1)$$

We used six categories defined by the Kidney Disease: Improving Global Outcomes (KDIGO) organization [20]. In the six categories, we only used data from participants whose category in the input year is G1, G2, and G3a (*i.e.* eGFR $\geqq 45\,\mathrm{mL/min/1.73\,m^2}$).

For attributes taking continuous values, we replaced missing values with the average value. For attributes taking discrete values, a missing value is treated as an additional category. These methods are commonly used to handle missing values in data used for training models.

Contradictory samples were removed from the original dataset. Specifically, we removed the participants who answered *no* to the question *Q2: Using insulin injection or antidiabetic (hypoglycemic) drug* in the output year, despite their stage being in *under medical treatment* that year. We assumed they did not answer the questions correctly and removed them from the dataset.

For HbA1c prediction using preprocessed data, the numbers of participants by time intervals were 1,410 for four years, 3,159 for three years, 5,281 for two years, and 7,811 for one year. For creatinine prediction, the numbers of participants by time intervals were 1,544 for four years, 3,466 for three years, 5,753 for two years, and 8,540 for one year.

We conducted five-times-five nested cross-validation to compare machine learning techniques. Namely, we divided the dataset into five folds. For each training session, we used one fold as a test dataset and the rest for training and validation. The folds not used for testing were split into five further folds. Four of them were used for training, and one was used for validation, that is, hyperparameter optimization. The dataset was split into folds participant-wise. In other words, no participant is contained in two or more folds.

4.4 Training and interpretation

We trained XGBoost [3] to predict the target attributes. It is known that XGBoost shows high performance for tabular data.

For interpretation, we used SHAP, which is based on game theory, to measure how each attribute contributed as a part of a coalition with other features in making the prediction correct. We used TreeExplainer [12] to calculate the SHAP values. There are model-independent methods for calculating SHAP values, such as Kernel SHAP. However, these methods compute SHAP values by making many predictions using perturbed input data. Therefore, the combinations become vast as the number of attributes increases and the computation time becomes longer. In TreeExplainer, SHAP values are calculated on the basis of the branching information used for prediction by the tree-based model, enabling fast and accurate calculation.

We optimized model parameters using the Optuna framework [1]. For each condition, we repeated training for 100 trials to optimize hyperparameters. In optimization, we changed the learning rate, max depth, min child weight, gamma, colsample by tree, and subsample as hyperparameter in the specified range. the learning rate is selected from {0.1, 0.01, 0.001}. max depth is selected from {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. min child weight is selected from {1, 2, 3, 4, 5}. gamma is selected from {0.0, 0.1, 0.2, 0.3, 0.4}. colsample by tree is selected from { 0.6, 0.7, 0.8, 0.9, 1.0} subsample is selected from {0.6, 0.7, 0.8, 0.9, 1.0}. We trained XGBoost for 1,000 rounds in each trial. For each method, if the validation root mean squared error (RMSE) did not improve for 20 rounds, we stopped training. After training, we selected the model having the highest validation RMSE and compared the results.

5 Evaluation

5.1 Prediction accuracy

Table 2 shows the RMSE, mean absolute error (MAE), and R^2 score for conditions where (1) the target attribute in an earlier year is used as one of the features and (2) the target attribute in an earlier year is not used as a feature. When the time interval was longer, all error measures increased. This indicates that the longer the time interval, the more difficult it is to make predictions. The graphs also show that errors increase when the target attribute is not used as a feature.

Table 2. Prediction performance of each condition

target	target attribute	time interval (year)	RMSE	MAE	R^2 score
HbA1c	included	1	0.188 ± 0.006	0.140 ± 0.003	0.709 ± 0.016
		2	0.200 ± 0.014	0.144 ± 0.004	0.685 ± 0.030
		3	0.233 ± 0.014	0.166 ± 0.005	0.614 ± 0.025
		4	0.234 ± 0.027	0.159 ± 0.011	0.603 ± 0.063
	not included	1	0.294 ± 0.007	0.226 ± 0.005	0.284 ± 0.029
		2	0.304 ± 0.015	0.230 ± 0.007	0.268 ± 0.039
		3	0.328 ± 0.019	0.245 ± 0.010	0.232 ± 0.020
		4	0.335 ± 0.023	0.245 ± 0.010	0.188 ± 0.052
Creatinine	included	1	0.055 ± 0.001	0.041 ± 0.001	0.876 ± 0.005
		2	0.059 ± 0.001	0.044 ± 0.001	0.858 ± 0.011
		3	0.062 ± 0.003	0.046 ± 0.002	0.847 ± 0.013
		4	0.065 ± 0.003	0.048 ± 0.001	0.838 ± 0.005
	not included	1	0.103 ± 0.003	0.080 ± 0.001	0.566 ± 0.007
		2	0.104 ± 0.003	0.081 ± 0.002	0.560 ± 0.027
		3	0.106 ± 0.003	0.082 ± 0.002	0.551 ± 0.013
		4	0.107 ± 0.005	0.082 ± 0.002	0.559 ± 0.024

5.2 HbA1c included as a feature

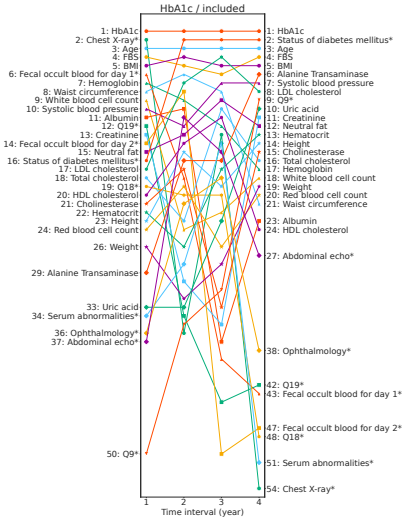
We first present the results for when HbA1c from an earlier year is included as a feature. For example, the amount of HbA1c in 2016 was used to predict HbA1c in 2020. Because there is a strong correlation between the amounts of HbA1c measured at two different years, the precision of prediction tends to be much higher than when not including them as a feature. Such high precision is useful for practical applications. However, because many predictions are explained by the amount of HbA1c in the earlier year, it is more difficult to see how other attributes contribute to making predictions. For this reason, including HbA1c as a feature might not be an ideal approach for scientific investigation on clarifying how various attributes affect the aggravation of the disease. Therefore, we also trained models in which HbA1c from an earlier year was not included as a feature. Such a model results in a lower prediction accuracy but enables to see contributing attributes other than HbA1c.

Figure 2a indicates how highly-ranked attributes change over a four-year period. The attributes are sorted in decreasing order of SHAP. The lines connect the same attributes across the years. The top three attributes (HbA1c, status of diabetes mellitus, and age) only slightly changed. However, attributes with lower ranks changed drastically over time. For example, waist circumference contributes largely to making predictions for the one-year interval (ranked 8th), but not much for the four-year interval (ranked 21st). In addition, alanine transaminase does not contribute for a short time interval (ranked 29th for the one-year interval) but contributes more in a longer time interval (ranked 6th for the four-year interval).

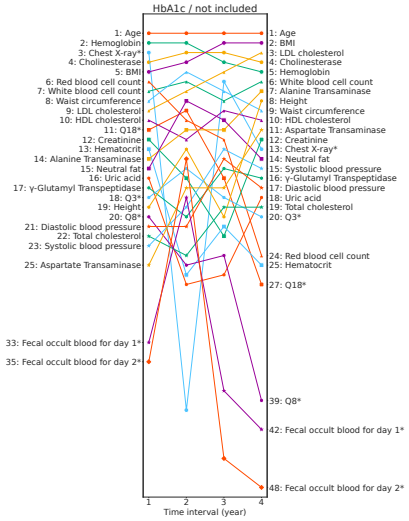
Figure 3a and 3b indicate the waist circumference and alanine transaminase for each sample, and the SHAP values corresponding to these values.

Alanine transaminase is one of the leading indicators of liver conditions. When the liver is in a normal condition, it works as an enzyme. When the liver's condition deteriorates, alanine transaminase, working inside the cells, leaks into the bloodstream, and its amount in the blood increases. Changes in the amount of alanine transaminase in the blood have been reported to be associated with diabetes [7, 22].

Among highly contributing attributes are uric acid, Q18, and Q19. Uric acid is one of the commonly used indicators of kidney function. The amount of uric acid in the blood increases when the kidneys are unable to filter it out. Causes for such deficiency are alcohol consumption or decreased kidney function. Q18 and Q19 are questions about the frequency and amount of alcohol consumption. Excessive drinking places an undue burden on the liver's ability to break down alcohol, resulting in a decline in liver function. As the time interval increases, the rank of uric acid goes up, while those of Q18 and Q19 go down. The fact that alanine transaminase and uric acid are more contributing in long-term predictions than in short-term ones may indicate that liver and kidney conditions have a long-term effect on diabetes. At the time of writing, we are unsure as to why the rankings of Q18 and Q19 go down over time. It could be because drinking habits may change over four years while uric acid stays the same.

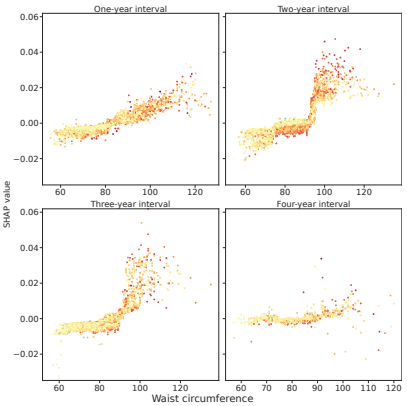


(a) The target attribute from an earlier year is included as a feature.

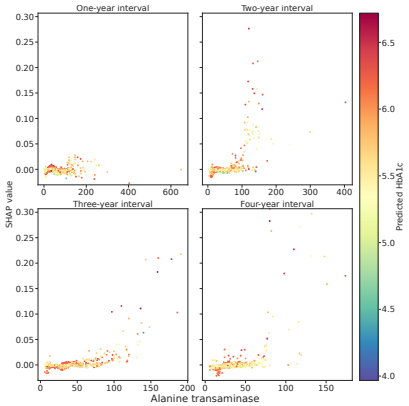


(b) The target attribute from an earlier year is **not included** as a feature.

Fig. 2. Ranking of attributes by SHAP values when predicting HbA1c.



(a) Waist circumference



(b) Alanine transaminase

Fig. 3. Waist circumference, alanine transaminase and SHAP values for each sample when predicting HbA1c using related attribute from earlier years.

5.3 HbA1c not included as a feature

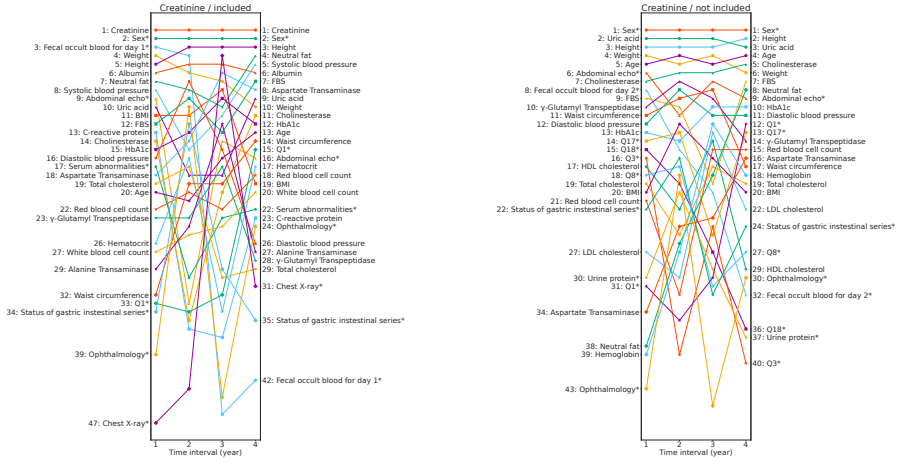
Figure 2b indicates the changes in high-ranked attributes. The change in ranking was nearly the same as the previous results when HbA1c was included as a feature. One difference is that hemoglobin appears as a high-ranked attribute, possibly due to its correlation to HbA1c.

For all time intervals, age was ranked highest. Cholinesterase, aspartate transaminase, and γ-glutamyl transpeptidase followed. It was essentially different from when HbA1c was used as a feature. Cholinesterase is an enzyme produced by the liver and is one of the indicators of liver function. It is highly correlated with nutritional status. The fact that diabetes is closely related to diseases such as a fatty liver may explain why cholinesterase is ranked high.

The attributes that changed their ranks significantly were Q3, Q8, and LDL cholesterol. Q3 is a question about using cholesterol-lowering drugs, and Q8 is a question about smoking. Cholesterol-related indices such as Q3 and LDL cholesterol have a significant relationship with the resultant condition of diabetes. Aspartate transaminase and γ-glutamyl transpeptidase are enzymes that work in the liver. These indices also fluctuate in value depending on the abnormalities of the liver.

These results suggest that attributes related to liver function contribute to making predictions of HbA1c. This coincides with our knowledge that diabetes strongly correlates with liver function and sugar in the blood.

5.4 Creatinine included as a feature



(a) The target attribute from an earlier year is included as a feature.

(b) The target attribute from an earlier year is **not included** as a feature.

Fig. 4. Ranking of attributes by SHAP values when predicting creatinine.

Figure 4a indicates how highly-ranked attributes change over a four-year period. The attributes are sorted in decreasing order of SHAP. The lines connect the same attributes across the years. As expected, creatinine is the top-ranked attribute. The other contributing attributes were sex, height, weight, albumin, neutral fat, and FBS. Creatinine is a substance produced by muscles throughout the body. Its amount increases or decreases depending on muscle mass. It is logical that sex, height, and weight, which affect muscle mass, are ranked high. Albumin is a type of protein found in the blood. Because it is produced in the liver, it represents how well the liver is functioning. If the liver is not working correctly, the production of albumin decreases. The kidneys filter out albumin, but if they are not functioning right, they may not be filtered out and run off into the urine.

There are attributes with significant changes. For example, fecal occult blood for the day is ranked 3rd in the one-year interval and 4th in the two-year interval. However, it is ranked below 40th for the three-year and four-year intervals. In addition, waist circumference goes up from being ranked 32nd (one-year interval) to 14th (four-year interval), suggesting its long-term effect on CKD.

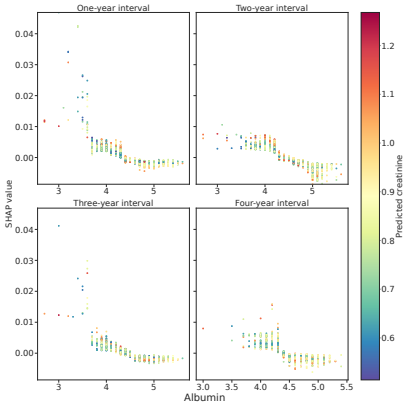


Fig. 5. Albumin and SHAP values for each sample when predicting creatinine using creatinine from earlier years.

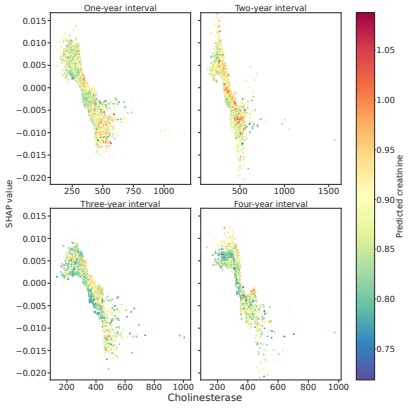


Fig. 6. Cholinesterase and SHAP values for each sample when predicting creatinine **without** using creatinine from earlier years.

Figure 5 shows that the SHAP value is higher in the positive direction when the albumin level is small. In particular, when kidney function deteriorates, nephrotic syndrome develops, which is a disease in which albumin in the blood flows out into the urine, decreasing in the amount in the blood. Therefore, the amount of albumin in the blood is as essential as creatinine when detecting the changes in kidney function at an early stage.

Waist circumference is also known as abdomen circumference, which increases due to the accumulation of fat in the gut and under the stomach's skin as the time

interval between features and the target attribute lengthens. The ranks of weight and BMI lower while that of waist circumference rises. Therefore, although one can estimate the degree of obesity from other indices, waist circumference is considered a clear indicator of fat accumulation and can represent the degree of obesity. In contrast, it cannot be defined only by height and weight.

5.5 Creatinine not included as a feature

Figure 4b shows the change of high-ranked attributes. When creatinine was not included as a feature, gender, height, and weight were ranked similarly as when creatinine was included. The other highly-ranked attributes were uric acid, cholinesterase, and abdominal echo. Uric acid and Cholinesterase are indicators of the function of the liver. Uric acid is also used to represent the level of kidney healthiness. Like HbA1c, they are also closely related to kidney function. Abdominal echo is an ultrasound examination of the abdominal organs such as the kidneys, liver, and pancreas. Therefore, it is reasonable that these attributes indicate the status of organs related to renal function and are listed as an essential attribute in the condition that does not use creatinine to make predictions. In Figure 6, when cholinesterase is large, the SHAP value decreases. Cholinesterase increases when the liver condition worsens, for example by having a fatty liver which is strongly associated with obesity. In addition, there is a strong relationship between obesity and a decrease in total body muscle mass. Therefore, it is reasonable that cholinesterase has a strong negative correlation with the SHAP value.

Among medical consultation questions, highly-ranked ones were Q1 and Q17. Q1 asks about the use of medication to lower blood pressure. Q17 asks about skipping breakfast three or more times a week. High blood pressure is caused by irregular sleep and disordered eating habits. Hypertension has a strong effect on blood vessels and has a significant impact on the kidneys. Therefore, it is logical that blood pressure status is substantial and contributes to the prediction of the kidney condition. The fact that diastolic blood pressure was ranked high also supports this effect.

A healthy diet is one of the essential factors in maintaining good health, regardless of kidney condition. In addition, breakfast provides the energy needed for daytime activities and moderates blood sugar fluctuations during the day. It also indicates the relationship between lifestyle and diet in maintaining kidney function.

6 Conclusion

We analyzed the time dynamics of relationships between the predictions in our predictive model and the attributes in health screening data. Overall, the combination of XGBoost and SHAP turned out to be extremely powerful for finding contributing attributes from health screening data.

The experiments showed that as the time interval between features and the target attribute changes, many attributes change their degree of significance. A number of them matched with our existing knowledge on the mechanism of diabetes and CKD, but there were also interesting, unexpected observations that may provide insight to medical researchers. For example, the rank of alanine transaminase rises as the time interval lengthens, both for predicting HbA1c and creatinine. This suggests that alanine transaminase is a good early indicator for the target diseases.

The investigation of time dynamics of interpretations can lead to finding new relationships between health screening and the progression of diabetes and CKD. In many cases, the results matched our knowledge regarding diabetes and CKD, suggesting the effectiveness of using interpretable machine learning to investigate the underlying mechanisms of diseases.

In this work, we interpreted models that predict future medical states using health records from a single year as an input. Our current approach cannot capture how the dynamics over several years affect the medical condition in the future. For example, our predictive model cannot take into account whether the patient's medical test result is deteriorating rapidly in a few years or not. If we can train predictive models that take health records from several years as input, we can see the effect of such dynamics on the future outcome. We, therefore, plan to develop predictive models that take time-series data as input and obtain interpretations for those models.

Acknowledgements

This work was supported by JST COI Grant Number JPMJCE1301 to Y.W.; by G-7 Scholarship Foundation, Uehara Memorial Foundation, and JSPS KAKENHI Grant Number 18KK0308 to T.T. We are grateful for the staff of Mito Kyodo General Hospital for data preparation and research support.

References

1. Akiba, T., et al.: Optuna: A next-generation hyperparameter optimization framework. In: Proc. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining. pp. 2623–2631 (2019)
2. Belur Nagaraj, S., et al.: Machine-learning–based early prediction of end-stage renal disease in patients with diabetic kidney disease using clinical trials data. Diabetes, Obesity and Metabolism **22**(12), 2479–2486 (2020)
3. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining. pp. 785–794 (2016)
4. Choi, B.G., et al.: Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. Yonsei Medical Journal pp. 191–199 (2019)
5. Dagliati, A., et al.: Machine learning methods to predict diabetes complications. Journal of Diabetes Science and Technology **12**, 295–302 (2018)

6. Dankwa-Mullan, I., et al.: Transforming diabetes care through artificial intelligence: The future is here. Population Health Man. **22(3)**, 229–242 (2019)
7. Itabashi, F., et al.: Combined associations of liver enzymes and obesity with diabetes mellitus prevalence: The tohoku medical megabank community-based cohort study. Journal of Epidemiology p. JE20200384 (2020)
8. Japanese Society of Nephrology: Clinical practice guidebook for diagnosis and treatment of chronic kidney disease 2012. The Japanese journal of nephrology **54**(8), 1031–1191 (2012)
9. Kunwar, V., et al.: Chronic kidney disease analysis using data mining classification techniques. In: 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence). pp. 300–305. IEEE (2016)
10. Lai, H., et al.: Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders **101**, 1–9 (2019)
11. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proc. of the 31st Int. Conf. on Neural Information Processing Systems. p. 4768–4777 (2017)
12. Lundberg, S.M., et al.: From local explanations to global understanding with explainable AI for trees. Nature Machine Intell. **2**(1), 56–67 (2020)
13. Marini, S., et al.: A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes. J. Biomedical Informatics (2015)
14. Ministry of Health, Labour and Welfare, Japan: Standardized questionnaire for health checkups (2013)
15. Moreno-Sanchez, P.A.: Features importance to improve interpretability of chronic kidney disease early diagnosis. In: 2020 IEEE Int. Conf. on Big Data (Big Data). pp. 3786–3792 (2020)
16. Park, S., et al.: Interpretable prediction of vascular diseases from electronic health records via deep attention networks. In: Proc. of IEEE 18th Int. Conf. on Bioinformatics and Bioengineering (2018)
17. Ravizza, S., et al.: Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nature medicine **25**(1), 57–59 (2019)
18. Shakeri, E., et al.: Exploring features contributing to the early prediction of sepsis using machine learning. In: 2021 43rd Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 2472–2475 (2021)
19. Sisodia, D., Sisodia, D.S.: Prediction of diabetes using classification algorithms. Procedia Computer Science **132**, 1578–1585 (2018)
20. Stevens, P.E., Levin, A.: Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. Annals of internal medicine **158**(11), 825–830 (2013)
21. Tangri, N., et al.: A predictive model for progression of chronic kidney disease to kidney failure. Jama **305**(15), 1553–1559 (2011)
22. Vozarova, B., et al.: High alanine aminotransferase is associated with decreased hepatic insulin sensitivity and predicts the development of type 2 diabetes. diabetes **51**(6), 1889–1895 (2002)
23. Wang, W., Chakraborty, G., Chakraborty, B.: Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm. Applied Sciences **11**(1), 202 (2021)
24. Xie, Z., et al.: Building risk prediction models for type 2 diabetes using machine learning techniques. Preventing chronic disease **16**, 1–9 (2019)