


Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan

Tadao Ooka ,¹ Hisashi Johno,² Kazunori Nakamoto,³ Yoshioki Yoda,⁴ Hiroshi Yokomichi,¹ Zentarō Yamagata¹

To cite: Ooka T, Johno H, Nakamoto K, *et al*. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. *BMJ Nutrition, Prevention & Health* 2021;**4**:e000200. doi:10.1136/bmjnph-2020-000200

► Additional material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjnph-2020-000200>).

For numbered affiliations see end of article.

Correspondence to

Dr Tadao Ooka, Department of Health Sciences, University of Yamanashi Faculty of Medicine Graduate School of Medicine, Chuo, Yamanashi, Japan; tohoka@yamanashi.ac.jp

Received 5 November 2020
Revised 23 February 2021
Accepted 25 February 2021
Published Online First
11 March 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

ABSTRACT

Introduction Early intervention in type 2 diabetes can prevent exacerbation of insulin resistance. More effective interventions can be implemented by early and precise prediction of the change in glycated haemoglobin A1c (HbA1c). Artificial intelligence (AI), which has been introduced into various medical fields, may be useful in predicting changes in HbA1c. However, the inability to explain the predictive factors has been a problem in the use of deep learning, the leading AI technology. Therefore, we applied a highly interpretable AI method, random forest (RF), to large-scale health check-up data and examined whether there was an advantage over a conventional prediction model.

Research design and methods This study included a cumulative total of 42908 subjects not receiving treatment for diabetes with an HbA1c <6.5%. The objective variable was the change in HbA1c in the next year. Each prediction model was created with 51 health-check items and part of their change values from the previous year. We used two analytical methods to compare the predictive powers: RF as a new model and multivariate logistic regression (MLR) as a conventional model. We also created models excluding the change values to determine whether it positively affected the predictions. In addition, variable importance was calculated in the RF analysis, and standard regression coefficients were calculated in the MLR analysis to identify the predictors.

Results The RF model showed a higher predictive power for the change in HbA1c than MLR in all models. The RF model including change values showed the highest predictive power. In the RF prediction model, HbA1c, fasting blood glucose, body weight, alkaline phosphatase and platelet count were factors with high predictive power.

Conclusions Correct use of the RF method may enable highly accurate risk prediction for the change in HbA1c and may allow the identification of new diabetes risk predictors.

INTRODUCTION

The number of diabetes cases worldwide is estimated at 463 million and is predicted to increase to 700 million by 2045.¹ Type 2 diabetes accounts for 90% of all diabetes cases, and three in four cases occur in the working age population.² Blood glucose and glycated haemoglobin A1c (HbA1c) are

What this paper add

- The prediction model based on the Random Forest method was able to predict the change in HbA1c with higher accuracy than that obtained with the regression analysis.
- Random forests showed some clinically important predictors that were not shown in the approach by regression analysis.
- Our findings suggest that machine learning methods such as the Random Forest method may be effective in detecting type 2 diabetes at a very early stage by predicting future increases in HbA1c.

often measured at regular medical check-ups, but active guidance and intervention are rarely performed unless the measured values exceed the standard values. It is suggested that insulin resistance increases in type 2 diabetes 10 years before its onset, therefore, early intervention before onset is important.³ For early intervention, it is necessary to build a model that efficiently predicts future increases in HbA1c.

Currently, there are an increasing number of research studies applying artificial intelligence (AI) technologies.⁴ In the medical field, AI is used in many types of research, such as discovery of new disease phenotypes,⁵ accurate diagnosis⁶ and cost-effectiveness prediction.⁷ In the future, AI is expected to be applicable to a wide range of medical research areas, from public health to molecular biology.⁸ In the area of public health, it is especially important to predict diseases from existing medical data with high accuracy and to propose appropriate early interventions for each individual.⁹ An ideal preventive care system with individual intervention may be realised by using AI.

Deep learning¹⁰ is a machine learning method representing AI technology that has

produced good results with respect to discrimination of pathological images¹¹ and fundus images.¹² However, this technology is limited in tracing its predictions back to the key discriminative features (ie, the ‘Black Box Problem’¹³), and typically requires large amounts of data for analysis.¹⁴ On the other hand, the random forest (RF) method, a machine learning method included in AI technology proposed by Breiman,¹⁵ can show the importance of variables used in its predictions.¹⁶ Recent studies have shown good results in terms of predicting Alzheimer’s disease and identifying predictors with RF.¹⁷

Regression analysis is often used in disease risk prediction studies; in particular, logistic regression analysis has been used in the recent study¹⁸ to generate models of diabetes risk in Japan. In the present study, we applied the RF method to medical data to investigate the advantages of this method over the existing method. Specifically, we applied the RF method to data obtained from annual health check-ups conducted on residents in Japan. We compared this method with an existing method by creating a model to predict the increase of HbA1c, and we also examined variables that influenced successful predictions with the RF method.

MATERIALS AND METHODS

Study participants

A total of 168206 data *samples* from 64379 *people* who received annual health check-ups at Yamanashi Koseiren

Healthcare Centre (Yamanashi, Japan) during April 1999 to March 2009, were included in this study. This annual check-up was performed based on a legal requirement imposed by the Industrial Safety and Health Act in Japan. With the goal of predicting the diabetes risk in a given year by using the results of the previous two consecutive years, we extracted the data for a total of 44307 data *samples* from 13253 *people* who had received a health check-up for three consecutive years. (A single person might have received multiple series of three-consecutive-years health check-ups within the 10-year study period. In these contexts, a data *sample* refers to a single series of three consecutive years health check-up data of a single person; *people* indicate the total number of the single persons who may have multiple data *samples* (for details, see online supplemental eMethod S1)

To ensure sufficient data for analysis, all data *samples* obtained from the same *people* were used redundantly in the analyses. *People* with HbA1c $\geq 6.5\%$ (48 mmol/mol) or those taking medicines for diabetes in either the first or second year were excluded from the analysis, yielding 42908 data *samples* (from 12977 *people*). Among these data *samples*, 32181 data *samples* (from 10408 *people*, 75% of data *samples*) were used as training data to develop each prediction model. We used 10727 data *samples* (from 8556 *people*, 25% of analysis data *samples*) as test data to draw receiver operating characteristic (ROC) curves for each model. We randomly extracted training and test

Table 1 Characteristics of the study participants

	(Unit)	All	Amount of HbA1c increase in the next year*						
			<0	0–0.1	0.2–0.3	0.4–0.5	0.6–0.7	0.8–0.9	1.0–
N		42908	13459	15761	9945	2898	536	158	151
Ggender†	%	53.10	53.51	52.20	52.64	54.07	60.63	71.52	74.83
Age	–	55.05	55.00	54.78	55.33	55.58	56.32	54.79	55.53
Height	Cm	161.71	161.64	161.77	161.60	161.66	162.57	163.35	164.68
Drink†	%	46.00	46.92	45.66	44.64	46.52	49.72	56.41	55.78
Smoke†	%	45.82	46.04	44.87	45.68	47.13	53.79	60.65	67.35
Weight	kg	60.08	59.99	59.86	60.00	60.94	61.82	65.14	68.13
BMI	–	22.87	22.85	22.77	22.87	23.21	23.32	24.28	24.96
Body fat	%	24.19	24.19	24.10	24.22	24.58	24.30	24.53	24.92
GTP	U/L	36.70	36.12	36.02	36.96	38.80	44.47	51.11	58.94
HDL-C	mg/dL	57.64	58.10	57.85	57.28	56.70	55.49	52.78	49.96
LDL-C	mg/dL	124.31	124.69	124.20	124.25	124.06	122.05	119.15	123.44
FBG	mg/dL	98.42	98.43	97.51	98.36	100.35	104.96	113.41	120.47
HbA1c	%	5.03	5.13	5.00	4.96	4.97	5.12	5.44	5.66
S-BP	mm Hg	124.79	125.28	124.00	124.50	126.56	127.90	129.91	130.65
D-BP	mm Hg	77.31	77.57	76.94	77.23	77.93	78.21	79.77	80.36

The definition of each variable’s abbreviation can be seen in table 2. Other characteristics can also be found in online supplemental table S2.

*The lower limit of HbA1c increase is -2.0% and the upper limit of HbA1c increase is $+5.6\%$.

†Gender: prevalence of male, drink: prevalence of drinking more than twice a week, smoke: prevalence of current smoking habit.

BMI, body mass index; DBP, diastolic blood pressure; FBG, fasting blood glucose; GTP, glutamyl transpeptidase; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure.

Table 2 Variables used in random forest method and multiple logistic regression models

Objective variable				
Model 1 ($\geq 0\%$)	1	HbA1c increase of $\geq 0\%$ from previous year†	0	HbA1c increase of not $\geq 0\%$ from previous year‡
Model 2 ($\geq 0.2\%$)	1	HbA1c increase of $\geq 0.2\%$ from previous year†	0	HbA1c increase of not $\geq 0.2\%$ from previous year‡
Model 3 ($\geq 0.4\%$)	1	HbA1c increase of $\geq 0.4\%$ from previous year†	0	HbA1c increase of not $\geq 0.4\%$ from previous year‡
Model 4 ($\geq 0.6\%$)	1	HbA1c increase of $\geq 0.6\%$ from previous year†	0	HbA1c increase of not $\geq 0.6\%$ from previous year‡
Model 5 ($\geq 0.8\%$)	1	HbA1c increase of $\geq 0.8\%$ from previous year†	0	HbA1c increase of not $\geq 0.8\%$ from previous year‡
Model 6 ($\geq 1.0\%$)	1	HbA1c increase of $\geq 1.0\%$ from previous year†	0	HbA1c increase of not $\geq 1.0\%$ from previous year‡
Explanatory variables (97 variables in total)				
Single-year value and change from previous year (92 (46+46) in total)		Weight, BMI, body fat, WCC, RCC, Hb, Ht, MCV, MCH, MCHC, PLAT, TP, ALB, A/G, ChE, T-Bil, D-Bil, I-Bil, ALP, LAP, GTP, LDH, AST, ALT, BUN, CRE, UA, Na, K, Cl, Ca, CK, TG, TC, LDL-C, HDL-C, FBG, HbA1c, S-BP, D-BP, FVC, FEV1, P-FVC, P-FEV1, CRP, RF		
Only single-year value (five in total)		Gender, age, height, drink*, smoke*		

*Drink: drinking more than twice a week (1) or not (0), smoke: Having current smoking habit (1) or not (0).

†the upper limit of HbA1c increase is +5.6%.

‡the lower limit of HbA1c increase is -2.0%.

A/G, albumin/globulin ratio; ALB, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; BUN, blood urea nitrogen; Ca, calcium; ChE, cholinesterase; CK, creatine kinase; Cl, chloride; CRE, creatinine; CRP, C reactive protein; D-Bil, direct bilirubin; D-BP, diastolic blood pressure; FBG, fasting blood glucose; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; GTP, glutamyl transpeptidase; Hb, haemoglobin; HbA1c, glycated haemoglobin A1c; HDL-C, high-density lipoprotein cholesterol; Ht, haematocrit; I-Bil, Indirect bilirubin; LAP, Leucine aminopeptidase; LDH, Lactate dehydrogenase; LDL-C, low-density lipoprotein cholesterol; MCH, mean corpuscular haemoglobin; MCHC, mean corpuscular haemoglobin concentration; MCV, mean corpuscular volume; Na, sodium; P-FEV1, forced expiratory volume % in one 1s; P-FVC, forced vital capacity %; PLAT, platelet; RCC, red cell count; RF, rheumatoid factor; S-BP, systolic blood pressure; T-Bil, total bilirubin; TC, total cholesterol; TG, triglyceride; TP, total protein; UA, urinary acid; WCC, white cell count.

data from all data *samples* (42 908 data *samples* from 13 253 *people*). Hence, there were some data *samples* derived from the same *people* across training and test data. Eventually, there were a total of 5987 *people* who had data *samples* in both training data and test data (see online supplemental figure S1).

Measurements

In creating the prediction model, the increase in HbA1c was designated as the objective variable, rather than the occurrence of diabetes (HbA1c $\geq 6.5\%$). Considering that our model was designed for early disease prevention, we should make the model that can apply to a wider range of *people* including those with lower HbA1c values. Therefore, we set up the prediction models based on increasing HbA1c values; the objective variables were dichotomised, depending on whether HbA1c increased by $\geq 0\%$, $\geq 0.2\%$, $\geq 0.4\%$, $\geq 0.6\%$, $\geq 0.8\%$ or $\geq 1.0\%$ after 1 year. These six prediction models were created in each analysis method. (Characteristics of the study participants according to the amount of HbA1c increase can be seen in table 1).

We included 97 items of the health check-up that were considered candidate explanatory variables (table 2). The result was expressed as a numerical or categorical value for 51 items and the remaining 46 items described how

the characteristic had changed in value from the previous year (items for which a change would be meaningless were not used, such as age and height).

RF method

RF is a machine learning method proposed by Breiman¹⁵ and is based on the ‘decision tree’ method used for non-parametric classification and regression. The decision tree is a method of classifying data by dividing it according to the value of a specific variable, then repeating this division, such that the divided data group consists of objective variables of the same category.

In the RF method, a decision tree is created using randomly selected variables for a data set extracted by bootstrap sampling, and classification is performed based on the majority of the decisions. In addition, the contribution of each variable to data classification can be determined using the created decision trees, and the importance of each variable can be calculated.¹⁹ ROC curves of RF models can be generated by changing the cut-off majority ratio of the involved decision trees. All analyses were conducted using R, V.3.6.1 (R Foundation for Statistical Computing, Vienna, Austria). Before using the RF method, the parameters *n*tree, the number of decision trees to be used, and *m*try, the number of variables

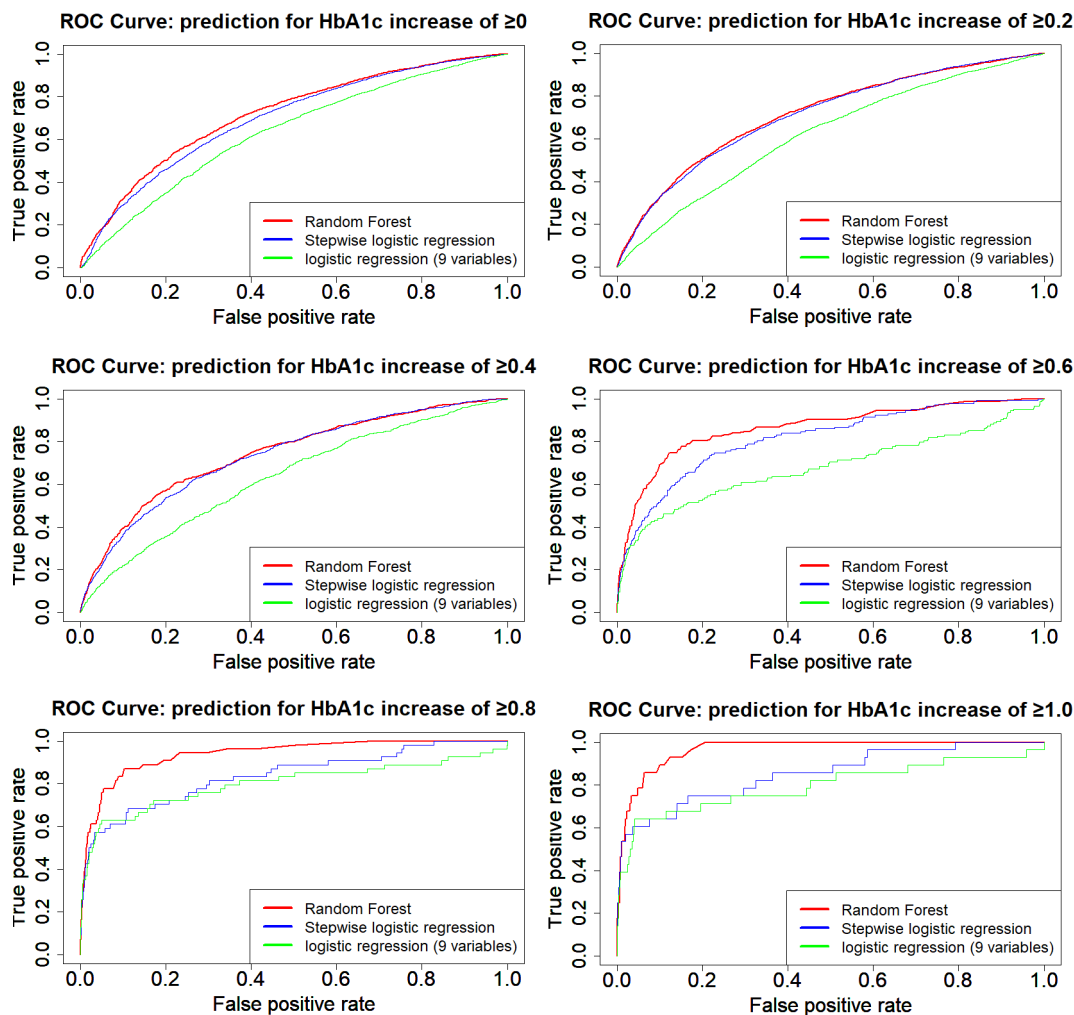


Figure 1 Receiver operating characteristic (ROC) curves showing the prediction performance of select models for changes in HbA1c. ROC curves of Random forest model (red line), multiple logistic regression model (stepwise logistic regression, blue line) and variable restricted multiple logistic regression model (logistic regression with nine variables, green line) are displayed according to the increase in HbA1c change value.

used to create the decision tree, need to be defined in advance. Parameters in the RF models were as follows: $n_{tree}=1000$, $m_{try}=9$ (RF model)/7 in variable-restricted RF (vrRF) model; all the other parameters were at default settings, on randomForest package V.4.6–14. The Gini index was used as an impurity function. For the detailed algorithm of the R packages used in this research, refer to a previous study by Biau and Scornet²⁰

Statistical analysis

In order to compare the performance of the machine learning model with the performance of the existing model, and to determine the contribution of the longitudinal data to the prediction accuracy, this study created four prediction models: an RF model, a multivariate logistic regression (MLR) model, a vrRF model and a variable-restricted MLR (vrMLR) model. In the RF model, all 97 items were used for prediction as explanatory variables. In the MLR model, stepwise analysis (using both forward and backward search with Akaike information criterion) was performed on 97 items for each of the six

types of objective variables (selected explanatory variables in each model are shown in online supplemental table S1). We restricted variables by excluding changes from the previous year and made predictions using 51 variables of a single year in the vrRF model. In the vrMLR model, predictions were restricted to only nine variables (not including the change value) that were used in recent studies of diabetes risk models in Japan.¹⁸ As the comparison of the ability of each method to select appropriate variables were included in this study, we did not select variables by taking into account previous studies, expert opinion or correlation coefficients of each variable in any method, except for the vrMLR model.

We constructed ROC curves of the four models, and each model was compared by calculating the area under the curve (AUC) for each ROC curve. In addition, the sensitivity and specificity at the optimum point, defined as the point of maximum value of the difference of the true positive rate and the false positive rate, of each ROC curve were calculated and compared. The variable importance

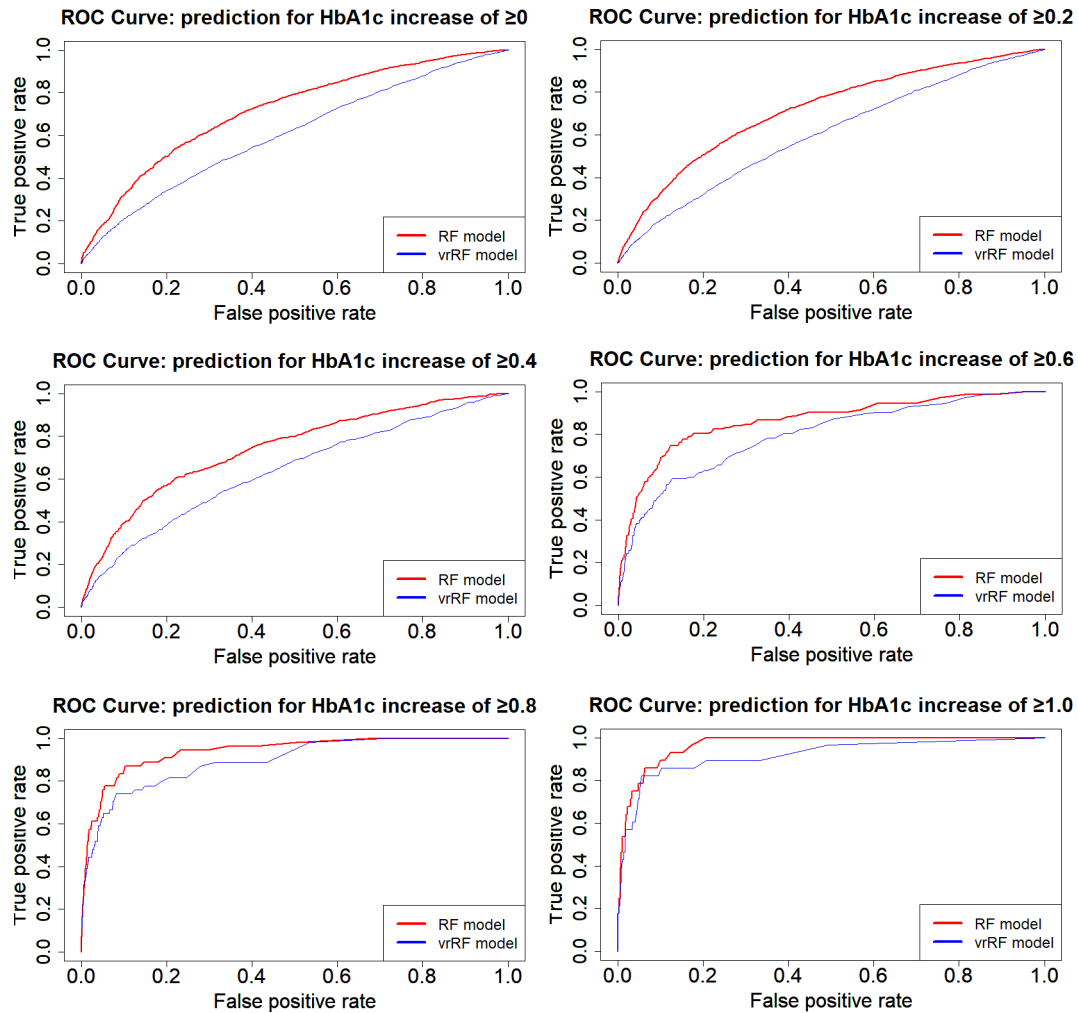


Figure 2 Receiver operating characteristic (ROC) curves showing the prediction performance of select models for changes in HbA1c. ROC curves of random forest model (using two consecutive years of values for prediction, red line) and variable restricted random forest model (using a single year of values for prediction, blue line) are displayed according to the increase in HbA1c change value.

(VI); defined in the prior literature,²¹ using the caret package on R, in the RF model and the standard regression coefficient (SRC) in the MLR model were also calculated. The 10 most important variables were enumerated in descending order of importance.

Patient and public partnership

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient-relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

RESULTS

Table 1 shows the descriptive statistics of representative variables associated with HbA1c, categorised according to the amount of HbA1c increase in the following year. An increasing trend in HbA1c was observed in several variables including gender, smoking, weight, body

mass index, glutamyl transpeptidase, fasting blood glucose (FBG), HbA1c and systolic blood pressure; a decreasing trend was observed in high density lipoprotein cholesterol.

Table 2 shows the objective variables and explanatory variables used in each model. For the RF method, all 97 variables were used as explanatory variables in all models. For the MLR method, conversely, different variables were adopted as explanatory variables in each model using the stepwise method. (For the explanatory variables selected in each model, please refer to online supplemental table S1).

Figure 1 shows the ROC curves for the RF, MLR and vrMLR models and figure 2 shows the ROC curves for the RF and vrRF models according to the increase in HbA1c change value, in order to compare prediction accuracy among models. Notably, the RF model showed the best predictive power among all models. By this comparison, we confirmed how changes from the previous year contribute to the prediction. In these ROC curves, the RF

Table 3 Sensitivity and specificity of best model and AUC on each ROC curve

	AUC	Best model on ROC curve			AUC	Best model on ROC curve	
		Sensitivity	Specificity			Sensitivity	Specificity
RF model1	0.719	0.714	0.617	MLR model1	0.699*	0.648	0.648
RF model2	0.716	0.608	0.720	MLR model2	0.711	0.648	0.668
RF model3	0.743	0.607	0.778	MLR model3	0.734	0.629	0.729
RF model4	0.864	0.804	0.823	MLR model4	0.817*	0.748	0.773
RF model5	0.940	0.870	0.898	MLR model5	0.840*	0.685	0.889
RF model6	0.967	0.929	0.877	MLR model6	0.854*	0.750	0.834
vrRF model1	0.606*	0.467	0.685	vrMLR model1	0.635*	0.610	0.605
vrRF model2	0.602*	0.516	0.632	vrMLR model2	0.622*	0.654	0.542
vrRF model3	0.638*	0.541	0.671	vrMLR model3	0.634*	0.607	0.594
vrRF model4	0.796*	0.594	0.874	vrMLR model4	0.680*	0.517	0.835
vrRF model5	0.895†	0.741	0.919	vrMLR model5	0.801*	0.630	0.950
vrRF model6	0.918	0.821	0.944	vrMLR model6	0.798*	0.643	0.959

*Significantly lower than that in the corresponding RF model: $p < 0.01$.

†Significantly lower than that in the corresponding RF model: $p < 0.05$.

AUC, area under the curve; MLR, Multiple Logistic Regression; RF, Random Forest; ROC, receiver operating characteristic; vrMLR, variable restricted multiple logistic regression (use only nine variables according to a previous study); vrRF, variable restricted random forest (only use single year for prediction).

model including the change value showed better prediction power than the vrRF model.

Table 3 shows the AUC for the ROC curve of each model, as well as the sensitivity and specificity at the optimum value on the ROC curve. All RF models had higher AUC than the other three types of models in all categories, and almost all had significant differences at a 5% level with some exceptions: MLR models 2 ($p=0.22$) and 3 ($p=0.31$) and the vrRF model 6 ($p=0.12$).

Table 4 shows the influence of each variable on HbA1c prediction. We set the degree of most influential variable in each model as 100% (SRC is converted to an absolute value because it can take a negative value). We extracted and compared the top 10 variables that showed the greatest contribution to the prediction using VI for the RF method and SRC for the MLR method. We also calculated the total rank of VI and SRC above all models by averaging the value of VI and SRC through all models.

DISCUSSION

In this study, we used a machine learning method, RF, to predict diabetes risk using HbA1c change values, and compared this model with MLR models. The results of this study suggest that RF-based models may have better performance for predicting changes in HbA1c than that of MLR-based models (table 3 and figure 1). Further, the RF models based on data of two consecutive years of health check-up may have better performance for predicting changes in HbA1c than that of the model based on data of only 1 year (figure 2). We also revealed that RF models used different factors from MLR models to make predictions (table 4). Therefore, we highlighted the significance of this machine learning method in

medical data analyses and presented diabetes predictive factors in a new format.

Many studies predicting disease risk have used regression analysis (ie, MLR or Cox proportional hazards regression). In contrast, the present study used the RF method, which showed better disease prediction accuracy than the existing model (MLR model). Therefore, the RF method may be more appropriate for suggesting predictors of disease risk than existing models under certain circumstances.

In RF models, the increase in HbA1c, HbA1c level, FBG level, the increase in FBG, and weight were selected as the most important variables. HbA1c, FBG, and weight were included in the diabetes risk models of previous studies.^{18 22} There have also been several previous studies to support the relationship between other items with high VI in the RF models and diabetes. For example, alkaline phosphatase has been suggested to be involved in glucose metabolism with other liver enzymes,²³ and platelet consumption have been suggested to increase in patients with diabetes.²⁴ C reactive protein has also been suggested to be strongly associated with the development of diabetes through the activation of adipocytes.²⁵

In the MLR model and in the 0%, 0.4%, 0.6% and 0.8% elevated HbA1c categories (models 1, 3, 4, 5), mean corpuscular haemoglobin (MCH), mean corpuscular volume (MCV) and MCH concentration (MCHC) were selected as the most important variables. In the 1.0% elevated HbA1c category (model 6), haematocrit, haemoglobin, and red cell count concentration were selected as the most important variables. The effect of these anaemia-related factors on HbA1c was supported in a previous study.²⁶

Table 4 Variable importance on random forest models and standard partial regression coefficient on multiple logistic regression models

RF model 1		RF model 2		RF model 3		RF model 4		RF model 5		RF model 6		Total		
Variables	VI*	Variables	VI*	Variables	VI*	Variables	VI*	Variables	VI*	Variables	VI*	Variables	VI*	
1	HbA1c_dif	100	HbA1c_dif	100	HbA1c_dif	100	HbA1c	100	HbA1c	100	HbA1c_dif	100	HbA1c_dif	100
2	HbA1c	66.8	HbA1c	41.6	FBG	53.3	HbA1c	96.6	HbA1c_dif	98.2	FBG	81.0	HbA1c	97.0
3	RF_dif	27.5	MCV_dif	28.7	HbA1c	50.8	FBG	87.2	FBG	81.9	HbA1c_dif	78.5	FBG	79.1
4	A/G_dif	27.0	FBG	28.4	FBG_dif	40.2	FBG_dif	65.0	FBG_dif	80.1	FBG_dif	62.7	FBG_dif	59.5
5	MCV_dif	25.2	RF_dif	25.7	MCV_dif	35.1	ALP	43.2	ALP	43.2	Weight	50.2	Weight	42.2
6	P-FEV1	23.3	MCHC_dif	25.1	CRP_dif	34.1	TC	43.1	Weight	41.7	ALP	42.9	ALP	41.2
7	P-FEV1_dif	23.2	FBG_dif	25.0	PLAT	33.3	weight	43.0	PLAT	40.0	TG	42.4	PLAT	38.7
8	GTP	22.7	P-FEV1_dif	25.0	FVC	33.3	CRP_dif	42.5	weight_dif	38.4	AST	40.8	CRP_dif	38.1
9	P-FVC_dif	22.4	P-FVC_dif	24.6	P-FEV1	32.6	PLAT	41.8	CK_dif	37.4	weight_dif	39.3	ALP_dif	38.0
10	CRP_dif	22.2	CRP_dif	24.4	ALP_dif	32.6	ChE	41.6	ALP_dif	37.2	ALP_dif	38.3	TG	37.7
MLR model 1		MLR model 2		MLR model 3		MLR model 4		MLR model 5		MLR model 6		Total		
Variables	SRC*	Variables	SRC*	Variables	SRC*	Variables	SRC*	Variables	SRC*	Variables	SRC*	Variables	SRC*	
1	MCH	100.0	HbA1c_dif	100.0	MCH	100.0	MCH	100.0	MCH	100.0	Ht	100.0	MCH	100.0
2	MCV	82.5	HbA1c	84.6	MCHC	53.2	MCV	61.9	MCV	64.7	Hb	80.0	MCV	72.6
3	MCHC	34.5	FBG	81.3	MCV	52.7	MCHC	46.2	MCHC	46.2	Hb_dif	32.9	MCHC	43.6
4	HbA1c	33.0	TC	75.7	Ht	46.5	RCC	30.0	FEV1	33.6	Ht_dif	30.9	Ht	43.2
5	Ht_dif	32.9	ALB_dif	71.4	RCC	43.7	Ht	29.1	MCH_dif	33.0	RCC	29.6	HbA1c_dif	39.1
6	TC	30.7	MCH	67.3	FVC	42.8	Hb_dif	24.6	FVC	29.5	MCHC	23.6	HbA1c	33.6
7	LDL-C	26.2	LDL-C	63.3	MCH_dif	41.5	RCC_dif	20.5	MCV_dif	28.3	MCV	18.8	FBG	32.6
8	HbA1c_dif	24.7	MCV	58.4	MCV_dif	39.4	MCV_dif	14.8	MCHC_dif	25.9	MCHC_dif	14.9	TC	29.5
9	FBG	20.6	HDL-C	51.1	HbA1c_dif	37.3	HbA1c_dif	12.8	Ht	24.4	HbA1c	8.2	RCC	27.1
10	HDL-C	19.4	A/G_dif	43.6	FBG	31.9	CI	10.6	RCC	23.1	TC	6.7	LDL-C	25.1

*Variable importance and standard partial regression coefficient are expressed as percentages. We set the degree of most influential variable in each model as 100% (SRC is converted to an absolute value because it can take a negative value).
 A/G, albumin/globulin ratio; CRP, C reactive protein; dif, Change value from the previous year; FBG, fasting blood glucose; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; GTP, glutamyl transpeptidase; Hb, haemoglobin; HbA1c, glycated haemoglobin A1c; HDL-C, high-density lipoprotein cholesterol; Ht, haematocrit; MCHC, mean corpuscular haemoglobin concentration; MLR, Multiple Logistic Regression; PLAT, platelet; RCC, red cell count; RF, Random Forest; SRC, Standard Regression Coefficient; TC, total cholesterol; VI, Variable Importance.

As mentioned above, both methods identified factors that had already been suggested to have associations with diabetes by previous studies. The value of VI in the RF model represents the ability of explanatory variables to clearly discriminate the group of each outcome (in this study, the increase level of HbA1c) by certain threshold level of the variables. Therefore, the VI can be used as a useful metric to identify the groups with exacerbations of type 2 diabetes in the early stages. Additionally, RF can adapt to the data with higher order interactions and non-linear effects, therefore, the VI have the ability to detect important variables for prediction even though there are non-linear relationships or strong interactions between the explanatory variables.²⁷ These properties of the VI may be the possible reasons of different model performance between RF and MLR models in this study.

Given the fact that each method identified different important factors, an approach that uses a variety of analytical methods should be considered when we intend to formulate a prediction model to identify important variables for prediction. For this kind of approach, the RF method is considered one of the appropriate analytical methods. Using the RF method, new disease predictors may be identified, such as more accurate prediction of various cancers, or new predictive factors of such cancers. Thus, the RF method can be regarded as an important analysis method in risk prediction research. However, it is necessary to confirm further applications and precise interpretation of this method.

A strength of this study is that it used several methods to predict the risk of diabetes for a large number of *people* who completed check-ups in the same facility. In addition, by designating the objective variable as the HbA1c change value, rather than the onset of diabetes, we were able to consider subjects whose HbA1c levels were originally low. Furthermore, we were able to examine which variables were strongly associated with diabetes prediction.

There were some limitations in this study. First, only those who received a medical check-up for three or more consecutive years were included for analysis. *People* who undergo many medical examinations may be more concerned about being healthy, so this predictive model may present a selection bias towards subjects with higher health consciousness. Second, because the presence or absence of treatment for diabetes was self-reported, those who falsely self-reported their health status may have been included as analysis subjects. These subjects are more likely to have elevated HbA1c because of missed medications, which may affect predictive models. However, it is important to identify these *people* because they are likely to experience a recurrence of diabetes. Third, we do not know whether the model will apply to other datasets because we did not confirm the external validity of the model created in this study. The model developed in this study can be applied only within the range of the training data and cannot predict outcomes with the values that fall outside the training data. Moreover, the investigation of important predictors revealed by the models could be

valid only to the populations similar to that of the training set (Japanese living in the countryside). To confirm external validity, it is necessary to confirm whether similar results are obtained in other health check-up facilities.

Given these limitations, the RF model predicted diabetes risk with significantly greater accuracy than existing models in the present study and identified highly relevant predictors. It is possibly beneficial to the medical field by utilising the RF method used in this research study. In terms of disease prevention, by incorporating the longitudinal data (such as continuous health check-up data) into this method, we can possibly predict the disease risk with higher accuracy than conventional risk models. In terms of disease treatment, we can identify which features of a patient are associated with serious outcomes by applying all data collected at admission to the RF model and comparing the VI on each feature.

However, we do not know whether this methodology will produce similar effects for outcomes other than HbA1c or from different types of datasets. Therefore, it is desirable to identify advantages of the RF method or other types of machine learning methods for diabetes or other disease prediction, through extrapolation to other medical data and validation of the present results.

Author affiliations

¹Department of Health Sciences, University of Yamanashi, Chuo, Yamanashi, Japan

²Department of Radiology, University of Yamanashi, Chuo, Yamanashi, Japan

³Center for Medical Education and Sciences, University of Yamanashi, Chuo, Yamanashi, Japan

⁴Yamanashi Koseiren Health Care Center, Kofu, Yamanashi, Japan

Acknowledgements This study was conducted in association with the Yamanashi Koseiren Healthcare Center. We are grateful to the staff members of the Yamanashi Koseiren Healthcare Center for careful data cleaning and to the Department of Health Sciences, University of Yamanashi, for careful proofreading of the manuscript.

Contributors TO designed the study. TO acquired data and analysed data. TO, HJ, KN and HY interpreted data. TO wrote the first draft of the manuscript. HJ, KN, HY, YY and ZY revised the manuscript critically. All authors have read and approved the final manuscript. TO is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding This work was supported by JSPS KAKENHI grant number 19K19433.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval All information used in this study was anonymised and analysed without identification of the individuals involved. The use of anonymised data for research is described on the website of the Yamanashi Koseiren Healthcare Center, and all subjects are given the opportunity to refuse participation in such analyses. This study was approved by the Research Ethics Committee of Faculty of Medicine, University of Yamanashi (receipt number: H30833).

Provenance and peer review Not commissioned; externally peer reviewed by 'Dr. Emmanuel Baah, University of North Carolina System, North Carolina, USA'.

Data availability statement The data used in this study are undisclosed data collected from health check-up recipients at the Yamanashi Koseiren Healthcare Center (Yamanashi, Japan) and are not publicly available. Yamanashi Koseiren Healthcare Center 1-1-26, Iida, Kofu, Yamanashi, Japan. Tel: +81 55-237-3630. <https://www.y-koseiren.jp/index.html>.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those

of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Tadao Ooka <http://orcid.org/0000-0002-1343-3986>

REFERENCES

- Saeedi P, Petersohn I, Salpea P, *et al*. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res Clin Pract* 2019;157:107843.
- Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. *Lancet* 2017;389:2239–51.
- Sagesaka H, Sato Y, Someya Y, *et al*. Type 2 diabetes: when does it start? *J Endocr Soc* 2018;2:476–84.
- Dean J, Patterson D, Young C. A new golden age in computer architecture: empowering the Machine-Learning revolution. *IEEE Micro* 2018;38:21–9.
- Shah SJ, Katz DH, Selvaraj S, *et al*. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* 2015;131:269–79.
- Krittawanong C, Zhang H, Wang Z, *et al*. Artificial Intelligence in Precision Cardiovascular Medicine. *J Am Coll Cardiol* 2017;69:2657–64.
- Chen P-J, Lin M-C, Lai M-J, *et al*. Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology* 2018;154:568–75.
- Zhu L, Zheng WJ. Informatics, data science, and artificial intelligence. *JAMA* 2018;320:1103–4.
- Dolley S. Big data's role in precision public health. *Front Public Health* 2018;6:68.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
- Ehteshami Bejnordi B, Veta M, Johannes van Diest P, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199–210.
- Ting DSW, Cheung CY-L, Lim G, *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017;318:2211–23.
- Fox M, Long D, Magazzeni D. Explainable planning. *arXiv* 2017.
- Liu B, Wei Y, Zhang Y. Deep neural networks for high dimension, low sample size data. *IJCAI* 2017:2287–93.
- Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- Qi Y. Random forest for bioinformatics. Ensemble machine learning. In: *Methods and applications*, 2012: 307–23.
- Lebedev AV, Westman E, Van Westen GJP, *et al*. Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *Neuroimage Clin* 2014;6:115–25.
- Nanri A, Nakagawa T, Kuwahara K, *et al*. Development of risk score for predicting 3-year incidence of type 2 diabetes: Japan epidemiology collaboration on occupational health study. *PLoS One* 2015;10:e0142779.
- Liaw A, Wiener M. Classification and regression by randomForest. *R news* 2002;2:18–22.
- Biau G, Scornet E. A random forest guided tour. *Test* 2016;25:197–227.
- Louppe G, Wehenkel L, Suter A. Understanding variable importances in forests of randomized trees. *Adv Neural Inf Process Syst* 2013:431–9.
- Noble D, Mathur R, Dent T, *et al*. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011;343:d7163.
- Cheung C-L, Tan KCB, Lam KSL, *et al*. The relationship between glucose metabolism, metabolic syndrome, and bone-specific alkaline phosphatase: a structural equation modeling approach. *J Clin Endocrinol Metab* 2013;98:3856–63.
- Michno A, Bielarczyk H, Pawelczyk T, *et al*. Alterations of adenine nucleotide metabolism and function of blood platelets in patients with diabetes. *Diabetes* 2007;56:462–7.
- Pradhan AD, Manson JE, Rifai N, *et al*. C-Reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA* 2001;286:327–34.
- Hong JW, Ku CR, Noh JH, *et al*. Association between the presence of iron deficiency anemia and hemoglobin A1c in Korean adults: the 2011-2012 Korea National health and nutrition examination survey. *Medicine* 2015;94:e825.
- Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat* 2007;1:519–37.