IEICE TRANS. INF. & SYST., VOL.E100–D, NO.8 AUGUST 2017

PAPER Special Section on Multiple-Valued Logic and VLSI Computing

Incidence Rate Prediction of Diabetes from Medical Checkup Data

Masakazu MORIMOTO[†a)], **Naotake KAMIURA**[†], **Yutaka HATA**[††], ***Members***, ***and*** **Ichiro YAMAMOTO**[†††], ***Nonmember***

SUMMARY To promote effective guidance by health checkup results, this paper predict a likelihood of developing lifestyle-related diseases from health check data. In this paper, we focus on the fluctuation of hemoglobin A1c (HbA1c) value, which deeply connected with diabetes onset. Here we predict incensement of HbA1c value and examine which kind of health checkup item has important role for HbA1c fluctuation. Our experimental results show that, when we classify the subjects according to their gender and triglyceride (TG) fluctuation value, we will effectively evaluate the risk of diabetes onset for each class.

key words: *specific health examination, lifestyle-related disease, machine learning*

1. Introduction

Japanese Ministry of Health, Labour and Welfare reported that the percentage of highly suspicious of diabetes is 15.5% for male and 9.8% for female in 2014 [1]. To decrease this percentage, we want to utilize individual annual health checkup reports to rise health consciousness. If we can predict a likelihood of developing lifestyle-related diseases from health check data, we will promote effective guidance by health checkup results. In this paper, we predict fluctuation of hemoglobin A1c (HbA1c) value, which deeply connected with diabetes onset. Here we focus on fluctuations in consecutive yearly health checkup data to investigate which element is related to HbA1c behavior.

To establish a framework to evaluate medical checkup data consisting on various domains, we first describe a fuzzy calculation method [2], [3]. We also analyze relationships between HbA1c and other items in specific health examination data, as one of the basic researches to establish the above measures, using self-organizing maps (SOMs) [4].

In this paper, to simplify the agenda, we are going to predict whether the HbA1c value exceed a threshold: 6.1 in JDS (Japan Diabetes Society) criterion or not, which corresponds to the worst health level: medical treatment and/or detailed examination. A set of specific health examination data is provided from Himeji Medical Association during the five-year period from 2008 through 2012. From the health checkup data, we can achieve 89427 subjects who has successive two-year data which isn't under treatment. Because this dataset has many missing value, we cannot achieve enough number of successive three-year data. In this dataset, 1569 subjects exceed the HbA1c threshold at second year. So its probability is only 1.75%. Figure 1 shows distribution of HbA1c value at 1st year, solid line represents the number of subjects who exceed the threshold in next year, and dashed line represents who did not exceed the threshold. From the figure, we can see that when a subject has enough small value at 1st-year, most of subjects doesn't exceed the HbA1c threshold in next year. So, we narrow the range of 1st year HbA1c value as $5.8 \leq \text{HbA1c} \leq 6.0$. In this condition, 1265 out of 9605 subjects exceeds the threshold in next year and its probability becomes 13.2%. For these dataset, we calculate yearly difference of following eleven health checkup data: body mass index (BMI), triglyceride (TG), high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), systolic blood pressure (SBP), gamma-glutamyl trans peptidase (GTP), creatinine (CRE), serum glutamic-oxaloacetic transaminase (GOT), serum glutamate pyruvate transaminase (GPT), uric acid (UA) and waist circumference, to predict the HbA1c value exceeds the threshold in second year.

There are many studies to predict developing diabetes from health checkup data by using machine learning techniques; Haffner et al. [5] use multivariate analysis, Wilson et al. [6] use regression model and Yu et al. [7] use support vector machine (SVM) to predict diabetes risk. In [8], Khalila et al. showed that Randomized Tree (RT) [9] based method outperformed other machine learning meth-

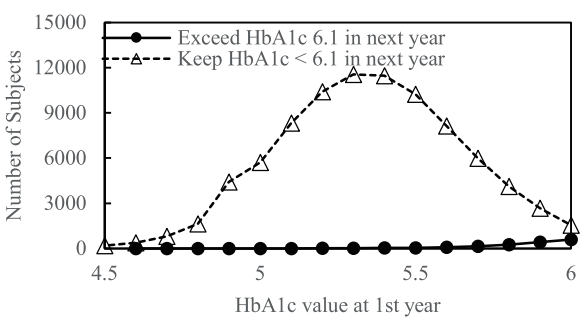


Fig. 1 Distribution of HbA1c.

Manuscript received October 14, 2016.
Manuscript revised February 23, 2017.
Manuscript publicized May 19, 2017.

[†]The authors are with the Graduate School of Engineering, University of Hyogo, Himeji-shi, 671–2280 Japan.
[††]The author is with the Graduate School of Simulation Studies, University of Hyogo, Kobe-shi, 650–0047 Japan.
[†††]The author is with Himeji Medical Association, Himeji-shi, 670–0061 Japan.
a) E-mail: morimoto@eng.u-hyogo.ac.jp

DOI: 10.1587/transinf.2016LOP0012

ods; SVM, bagging and boosting, in terms of the area under the receiver operating characteristic (ROC) curve (AUC).

In this paper, we also employ RT method to predict developing diabetes. However, our purpose is not further improvement of prediction accuracy but establishing effective report to rise health consciousness of individual subjects. By the analysis of prediction results, we can see that what kind of health checkup item has relation with HbA1c increment and we classify the type of diabetes onset. By evaluating the risk of developing diabetes for each class, we can make personal advice which propose numerical targets of health checkup items and show risk reduction by the achievement.

2. Predict HbA1c Increment from Fluctuation of the Other Health Checkup Items

To predict a subject's HbA1c value exceeds the threshold in 2nd year or not, we use RT method. The RT (sometimes called *random forest*) is an ensemble learning method for classification, that fits a number of decision tree (DT) classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. We use the eleven health checkup data as an input vector, and classified whether the HbA1c value exceed the threshold in second year or not. By analyzing the prediction result, we can see what kind of parameter will change with HbA1c.

2.1 Baseline Performance

First of all, by using RT, we test the prediction for 9605 subjects, whose HbA1c value was between 5.8 and 6.0 at 1st year. Figure 2 shows a prediction result as ROC curve. In this figure, horizontal axis represents false positive rate (FPR); the rate of healthy subjects who incorrectly receive a positive prediction result, and vertical axis represents true positive rate (TPR); the rate of unhealthy subjects who receive correct prediction result. The AUC value, which is a common criterion for evaluating ROC curve, becomes 0.67 for RT. This is a baseline performance of this prediction. From the preliminary experiments, we select RT parameters as follows, max depth is 4 and number of trees is 100.

Here, we also show results of other machine learning methods; DT and logistic regression model (LRM) in Fig. 2. From the results, RT slightly outperforms LRM (AUC = 0.66) and has advantage against to DT (AUC = 0.63).

2.2 Tendency by Gender

Next, we divide the health checkup data by gender, because our earlier study indicates that there exist several routes to diabetes onset, especially, sexual difference has strong influence to HbA1c increment [4]. In the dataset, we have 3636 male subjects and 5761 female subjects who has 1st year HbA1c value between 5.8 and 6.0. For each gender group, we apply RT to predict whether the subject exceeds HbA1c threshold in next year or not.

Figure 3 shows a prediction result by ROC curve. Here,

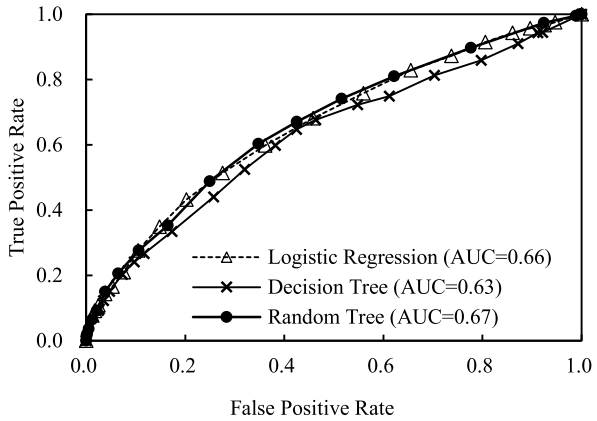


Fig. 2 Prediction results whether HbA1c value exceed the threshold from yearly difference of other eleven health checkup data.

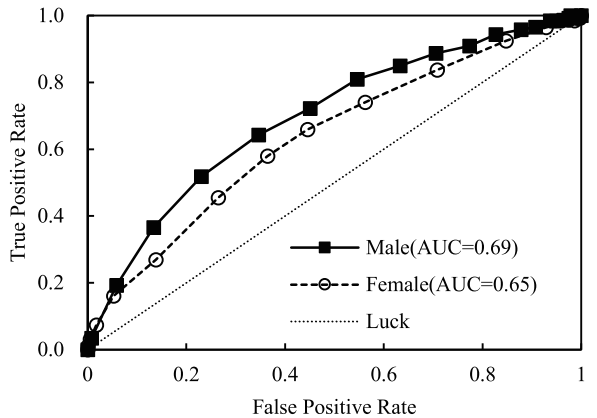


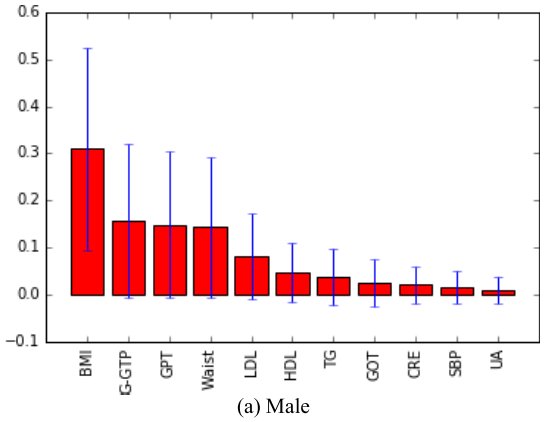
Fig. 3 Prediction results for each gender.

we select RT parameters as follows; number of trees = 100, max depth of tree = 3, and we apply 10-fold cross validation to make the ROC curve by changing weight parameter, which can control class priority. Here, we also show AUC value, which represent the prediction accuracy; 1 represents a perfect prediction and .5 represents a worthless prediction.

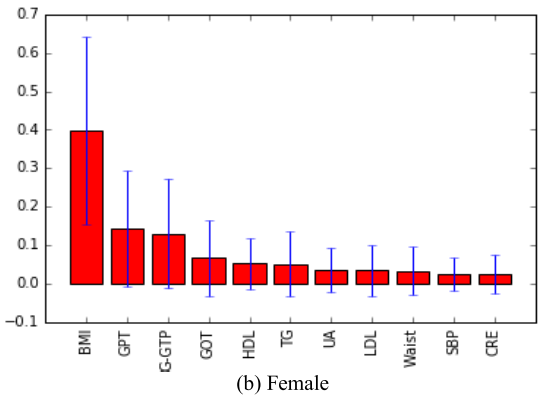
From this figure, we can see that the accuracy of male prediction is superior to female's one. This result implies that the relationship between HbA1c fluctuation and others will have simpler connection than female's one. To discuss this, we calculate feature importance of each health checkup item for both groups. Here, the importance is the Gini importance, which is computed as the total reduction of the criterion brought by that feature. Figure 4 (a) and 4 (b) shows the feature importance of RT classifier for male and female, respectively. From the result, BMI fluctuation is most important for predicting whether HbA1c excess the threshold, and second and third important item is GPT and GTP which represents liver function for both gender group. On the other hand, waist fluctuation is important only for male group.

2.3 Simple Grouping of Subjects

There will be several causes for increment of HbA1c, so it



(a) Male



(b) Female

Fig. 4 Feature importance of each health checkup item.

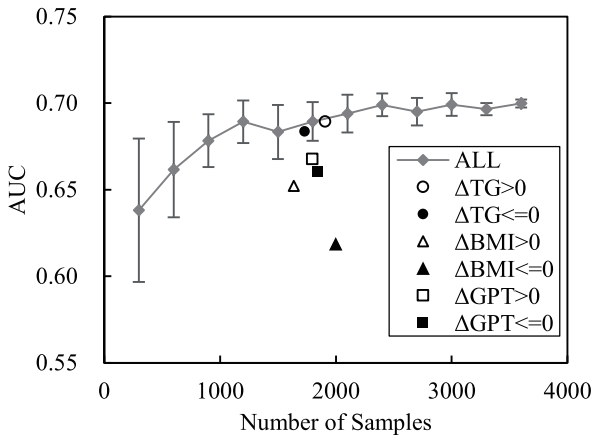
will be difficult to predict whole subjects by single predictor. To confirm this assumption, here we test several simple grouping of the subjects. In this section, we pick up three health checkup item, BMI, GPT and TG. As shown in Fig. 4, BMI and GPT have enough importance for predicting HbA1c increment, and our previous study [4] also indicate that TG has some sort of relation with HbA1c increment. Here we simply divide the dataset whether the item value was increased or decreased for each gender set.

Table 1 shows the number of subjects for each group and their prediction accuracy by AUC value. The increment of selected three items has adverse effect for health, so the increased groups have larger HbA1c excess ratio. Contrary to our expression, prediction accuracy drops by grouping at most cases. To separate adverse effect of reducing the number of training samples, we examine the relation between number of training samples and prediction accuracy. Figure 5 shows AUC comparison against to the number of training samples. From these results, we can see that grouping by BMI decreases prediction accuracy significantly. As shown in Fig. 5, BMI fluctuation is a most significant clue for prediction, but the dataset dividing by BMI fluctuation will lose the advantage and complicate the problem.

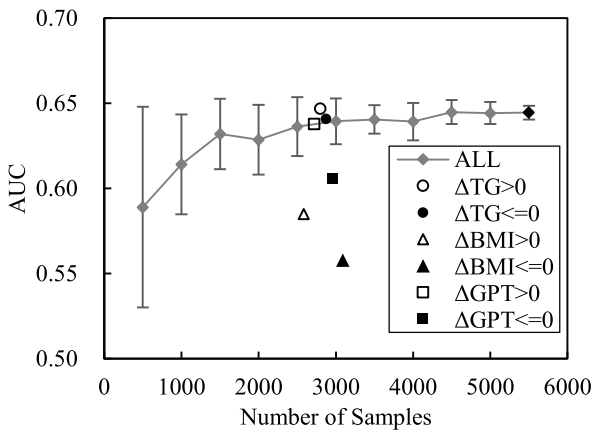
Figure 6 shows examples of DT for predicting whether HbA1c value exceeds threshold at next year or not. In this figure, TP represents true-positive, which can correctly predict the HbA1c exceedance, and FP, TN, FN are false-

Table 1 HbA1c prediction accuracy by grouping.

Gender	Group	Subjects	HbA1c excess		AUC
			number	ratio(%)	
Male	All	3636	552	15.2	0.698
	Δ BMI>0	1637	361	22.1	0.652
	Δ BMI<=0	1999	191	9.6	0.618
	Δ GPT>0	1796	359	20.0	0.668
	Δ GPT<=0	1840	193	10.5	0.661
	Δ TG>0	1906	330	17.3	0.689
	Δ TG<=0	1730	222	12.8	0.683
Female	All	5671	664	11.7	0.644
	Δ BMI>0	2584	428	16.6	0.585
	Δ BMI<=0	3087	236	7.6	0.558
	Δ GPT>0	2720	374	13.8	0.638
	Δ GPT<=0	2951	290	9.8	0.606
	Δ TG>0	2799	353	12.6	**0.647**
	Δ TG<=0	2872	311	10.8	0.641



(a) Male



(b) Female

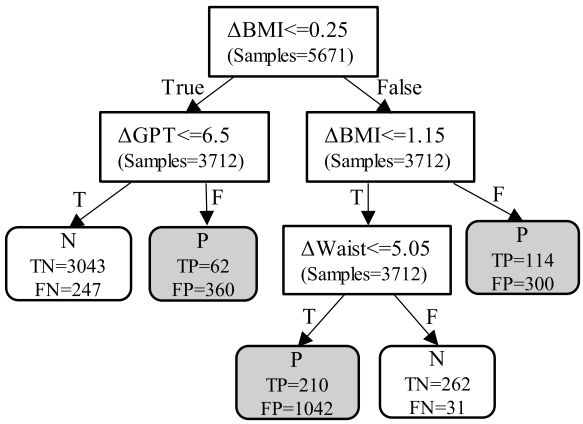
Fig. 5 Average AUC of grouping dataset according to the number of samples.

negative, true-negative and false-negative, respectively. Figure 6 (a) shows a tree for whole female dataset and Fig. 6 (b) shows a tree for female who increase her TG value in next year. From these figures we can confirm that BMI fluctuation plays important role for decision in each cases, and by grouping dataset, we can simplify the tree and improve its accuracy for HbA1c prediction.

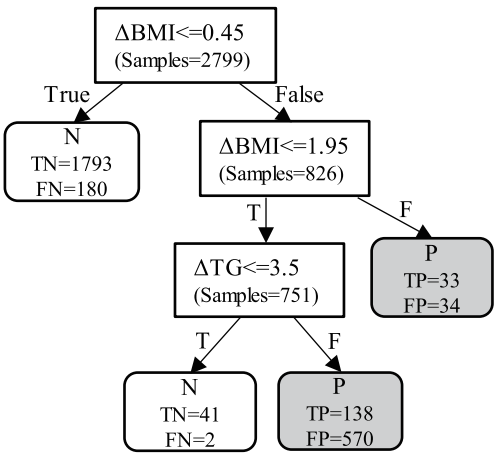
3. Incidence Rate Prediction by Using Decision Tree

In the previous section, we are going to predict whether the HbA1c value exceeds a threshold or not, by using fluctuation of other 11 health checkup items, but, its prediction accuracy is not so high. This is because that there are many causes to onset diabetes and subjects have wide individual differences. However, from the generated decision tree, we can estimate incidence rate of HbA1c exceedance in detail. It will be useful to rise health consciousness by showing specific guidance.

Figure 7 shows a decision tree to predict the HbA1c exceedance for female dataset whose HbA1c value of the first year was 5.8. The total exceedance risk of female, whose HbA1c value was 5.8, is 4.5%, but if the person increase BMI more than 0.95 and GTP more than 6.5 until next year, her exceedance risk becomes 27.9%. On the other hand, if she can keep BMI increment under 0.95, her exceedance risk falls below 3.5%. A specific guidance showing numerical target will enhance the motivation to avoid lifestyle-related diseases.



(a) A result tree for whole female subjects
(TPR=0.58, FPR=0.34, Accuracy=0.65)



(b) A result tree for $\Delta TG > 0$, female subjects
(TPR=0.48, FPR=0.25,Accuracy=0.71)

Fig. 6 Examples of Decision Tree for HbA1c prediction.

4. Conclusion

In this paper, we have tried to predict the HbA1c increment by a Machine Learning method from 11 health checkup data fluctuations. Some experimental result show that by grouping subject according to their TG fluctuation, we can slightly improve the prediction accuracy. We also show that by showing tree structure, we will make an effective health guidance by showing specific numerical target. In future study, we are going to improve prediction accuracy of diabetes incidence, by establishing effective classification of subjects.

Acknowledgments

This research was supported in part by Japan Society for the Promotion of Science with Grant-in-Aid for Scientific Research (A) (Grant number JP25240038).

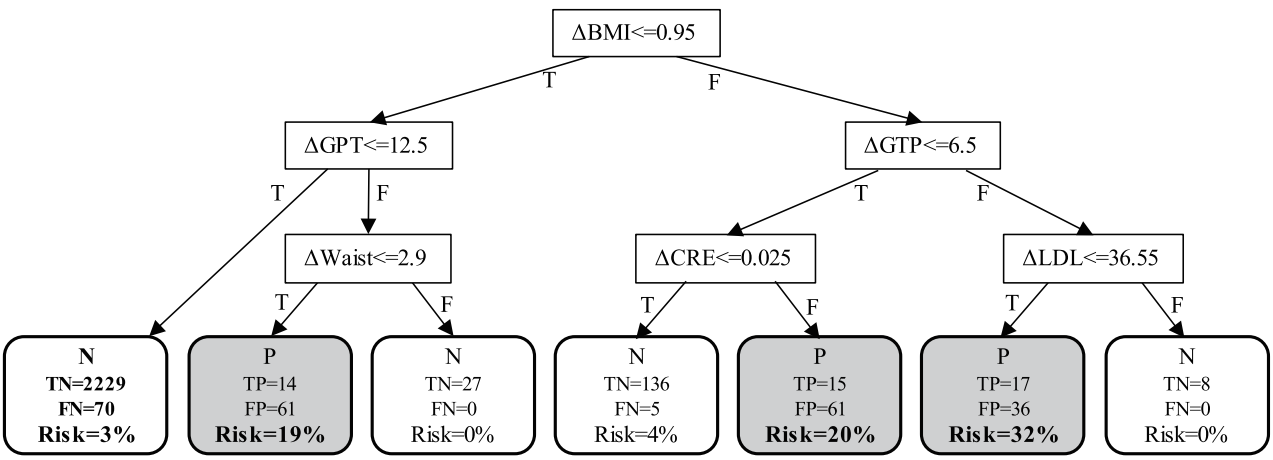


Fig. 7 An example of risk evaluation by Decision Tree for female with 1st year HbA1c = 5.8.

References

[1] Ministry of Health, Labour and Welfare's home page, http://www.mhlw.go.jp/
[2] S. Higuchi and Y. Hata, "Fuzzy Logic Approach to Health Checkup Data Analysis," Proc. 2014 World Automation Cong., pp.388–393, 2014.
[3] S. Higuchi and Y. Hata, "Fuzzy Dependency Analysis for Medical Checkup Reference," Proc. 2014 IEEE Int. Conf. on Systems, Man and Cybernetics, pp.4010–4015, 2014.
[4] H. Komori, S. Kobashi, N. Kamiura, Y. Hata, and K. Sorachi, "On relationship analysis of health examination items using self-organizing maps," Proc. 2015 Int. Conf. on Informatics Electronics & Vision (ICIEV), 2015, DOI: 10.1109/ICIEV.2015.7334004
[5] S.M. Haffner, M.P. Stern, B.D. Mitchell, H.P. Hazuda, and J.K. Patterson, "Incidence of type II diabetes in Mexican Americans predicted by fasting insulin and glucose levels, obesity, and body-fat distribution," Diabetes, vol.39, no.3, pp.283–288, 1990.
[6] P.W.F. Wilson, J.B. Meigs, L. Sullivan, C.S. Fox, D.M. Nathan, and R.B. D'Agostino, "Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study," Archives in Internal Medicine, vol.167, no.10, pp.1068–1074, 2007.
[7] Y. Wei, et al., "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," BMC Medical Informatics and Decision Making, 2010 10:16, DOI: 10.1186/1472-6947-10-16
[8] M. Khalila, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," BMC Medical Informatics and Decision Making, 2011 11:51, DOI:10.1186/1472-6947-11-51
[9] L. Breiman, "Random Forests," Machine Learning, vol.45, pp.5–32, 2001, DOI:10.1023/A:1010933404324

Masakazu Morimoto received B.E., M.E. and D.E. degrees in communication engineering from Osaka University in 1992, 1996 and 1998, respectively. Since 1998, He has been with Himeji Institute of Technology (currently University of Hyogo), Himeji, Japan, where he is an associate professor of the Graduate School of Electrical Engineering. His research interests are in the area of image recognition and its applications.

Naotake Kamiura received the B.E. degree (Electronic Engineering) in 1990, the M.E. degree (Electronic Engineering) in 1992 and the D.E. degree (Doctor of Engineering) in 1995 from Himeji Institute of Technology, Japan. He is currently a professor in the Department of Electronics and Computer Science, Graduate School of Engineering, University of Hyogo. His research interests include the application of soft computing to medical engineering. He is members of the Japanese Society of Medical Imaging Technology and the IEEE.

Yutaka Hata received the B.E. degree in 1984, the M.E. degree in 1986 and the Ph.D, in 1989 all from Himeji Institute of Technology, Japan. He is currently a Professor in the Graduate School of Simulation Studies, University of Hyogo, Japan. He spent one year in BISC Group, University of California at Berkeley from 1995 to 1996 as a visiting scholar. His research interests are in medical system, health monitoring system, fuzzy system and Immune system. He received 13 international awards such as the Franklin V. Taylor Best Paper Award (IEEE SMC 2009), etc. He is 5 Journal editors including IEEE trans. SMC-Systems. He is an IEEE Fellow.

Ichiro Yamamoto received the M.D. and Ph.D. degrees from Kawasaki Medical University in 1983 and from Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University in 1987, respectively. He established Yamamoto Clinic of Internal and Gastroenterological Medicines in 1988, and has been engaged in community medicine for citizens of Himeji as the head-physician of his clinic. He is currently the president of Himeji Medical Association.