Original Paper

Machine Learning Approach to Decision Making for Insulin Initiation in Japanese Patients With Type 2 Diabetes (JDDM 58): Model Development and Validation Study

Kazuya Fujihara[1], MD, PhD; Yasuhiro Matsubayashi[1], MD, PhD; Mayuko Harada Yamada[1], MD, PhD; Masahiko Yamamoto[1], MD, PhD; Toshihiro Iizuka[2], MSc; Kosuke Miyamura[2], MSc; Yoshinori Hasegawa[2], LLB; Hiroshi Maegawa[3], MD, PhD; Satoru Kodama[1], MD, PhD; Tatsuya Yamazaki[4], PhD; Hirohito Sone[1], MD, PhD, FACP

[1]Department of Internal Medicine, Faculty of Medicine, Niigata University, Niigata, Japan

[2]NTT Comware Corporation, Tokyo, Japan

[3]Department of Internal Medicine, Shiga University of Medical Science, Shiga, Japan

[4]Faculty of Engineering, Niigata University, Niigata, Japan

Corresponding Author:
Hirohito Sone, MD, PhD, FACP
Department of Internal Medicine
Faculty of Medicine
Niigata University
1-757 Asahimachi-dori Chuoh-ku Niigata
Niigata, 9518510
Japan
Phone: 81 25 368 9026
Email: sone@med.niigata-u.ac.jp

Abstract

Background: Applications of machine learning for the early detection of diseases for which a clear-cut diagnostic gold standard exists have been evaluated. However, little is known about the usefulness of machine learning approaches in the decision-making process for decisions such as insulin initiation by diabetes specialists for which no absolute standards exist in clinical settings.

Objective: The objectives of this study were to examine the ability of machine learning models to predict insulin initiation by specialists and whether the machine learning approach could support decision making by general physicians for insulin initiation in patients with type 2 diabetes.

Methods: Data from patients prescribed hypoglycemic agents from December 2009 to March 2015 were extracted from diabetes specialists' registries, resulting in a sample size of 4860 patients who had received initial monotherapy with either insulin (n=293) or noninsulin (n=4567). Neural network output was insulin initiation ranging from 0 to 1 with a cutoff of >0.5 for the dichotomous classification. Accuracy, recall, and area under the receiver operating characteristic curve (AUC) were calculated to compare the ability of machine learning models to make decisions regarding insulin initiation to the decision-making ability of logistic regression and general physicians. By comparing the decision-making ability of machine learning and logistic regression to that of general physicians, 7 cases were chosen based on patient information as the gold standard based on the agreement of 8 of the 9 specialists.

Results: The AUCs, accuracy, and recall of logistic regression were higher than those of machine learning (AUCs of 0.89-0.90 for logistic regression versus 0.67-0.74 for machine learning). When the examination was limited to cases receiving insulin, discrimination by machine learning was similar to that of logistic regression analysis (recall of 0.05-0.68 for logistic regression versus 0.11-0.52 for machine learning). Accuracies of logistic regression, a machine learning model (downsampling ratio of 1:8), and general physicians were 0.80, 0.70, and 0.66, respectively, for 43 randomly selected cases. For the 7 gold standard cases, the accuracies of logistic regression and the machine learning model were 1.00 and 0.86, respectively, with a downsampling ratio of 1:8, which were higher than the accuracy of general physicians (ie, 0.43).

Conclusions: Although we found no superior performance of machine learning over logistic regression, machine learning had higher accuracy in prediction of insulin initiation than general physicians, defined by diabetes specialists' choice of the gold

standard. Further study is needed before the use of machine learning–based decision support systems for insulin initiation can be incorporated into clinical practice.

(JMIR Med Inform 2021;9(1):e22148) doi: 10.2196/22148

KEYWORDS

hypoglycemic prescription; diabetes specialists; initial therapy; patterns of usage; machine learning

Introduction

While oral antihyperglycemic agents are indicated for many patients with type 2 diabetes, some patients require insulin injections, with or without oral antihyperglycemic agents, in the advanced stages of diabetes. Since type 2 diabetes typically develops and progresses gradually and asymptomatically [1], it is often found at the first primary care consultation at a rather advanced stage with fatigue, thirst, and polyuria accompanied by substantially elevated plasma glucose levels. Such situations force physicians to judge whether to prescribe insulin as the initial therapy to avoid further disease progression. A physician's misjudgment sometimes results in a hyperglycemic coma or another serious condition, as most patients hesitate to use insulin therapy because of inconvenience and cost [1,2]. Since there are no absolute standards for judgment of insulin initiation, this important decision made at the first consultation in primary care must be based on the physician's knowledge of the pathophysiology of the patient's condition and much prior experience. While diabetes specialists, defined as board-certified diabetologists, are trained on whether to choose insulin therapy based on their perception of the existence of complex conditions in their patients, as well as their overall health [3-5], such judgments are not easy for nonspecialists, defined as general physicians without board certification as diabetologists.

Machine learning, which can learn patterns and decision rules from data [6-9], has been used in clinical practice. Applications of machine learning for the early detection of diabetic retinopathy and cancer, for which clear-cut diagnostic gold standards exist, have been evaluated [10-16]. However, little is known about the usefulness of machine learning for decisions such as insulin initiation by specialists, for which there are no absolute criteria for use in clinical settings.

In this study, we first evaluated the ability of machine learning models to predict insulin initiation by specialists using the Japan Diabetes Clinical Data Management (JDDM) Study Group, which consists of diabetes specialists. Then, we compared the clinical decisions made by the machine learning approach (trained using the database of specialists' judgments) with those made by nonspecialists regarding whether to prescribe insulin for patients with type 2 diabetes at the first consultation. Using this information, we attempted to clarify the ability of machine learning models and determine whether artificial intelligence might assist clinicians in deciding on the initial therapy for type 2 diabetes in clinical practice.

Methods

Study Participants

Data were extracted from patients prescribed hypoglycemic agents from December 2009 to March 2015 using software (CoDiC) developed by the JDDM Study Group to promote clinical research on diabetes. Details on the JDDM Study Group and CoDiC are described elsewhere [3,4,17,18]. Briefly, the JDDM Study Group is a large network of diabetes specialists in Japan in 98 facilities. Study participants were individuals aged 20 years or older who started medical treatment for type 2 diabetes in outpatient clinics. Of the 6864 participants who received initial monotherapy during the above time period, we excluded 2004 individuals because of missing data on covariates (age, sex, BMI, duration of diabetes, level of glycated hemoglobin [HbA_{1c}], hypertension, and estimated glomerular filtration rate [eGFR]). Thus, data were analyzed from 4860 patients who were prescribed antidiabetic medications including insulin as the initial medical treatment and had laboratory data (Table 1). The ethics committee of the JDDM Study Group and Niigata University approved this study (2012-7, 2017-0294). Informed consent was obtained from all patients at each participating institute in accordance with the Guidelines for Epidemiological Studies of the Ministry of Health, Labour and Welfare of Japan.

Table 1. Characteristics of study participants according to prescription of insulin or another hypoglycemic drug.

Characteristic	Insulin (n=293)	Noninsulin (n=4567)	*P* value
Age (years), mean (SD)	59 (14)	61 (13)	.009
Age (years), n (%)			
<40	31 (11)	256 (6)	<.001
40-59	114 (39)	1635 (36)	
≥60	148 (51)	2676 (59)	
Male-to-female ratio	195:98	2929:1638	.40
BMI (kg/m^2), mean (SD)	24.5 (4.2)	25.7 (4.6)	<.001
BMI (kg/m^2), n (%)			
<22.5	96 (33)	1045 (23)	<.001
22.5-25.0	81 (28)	1158 (25)	
≥25.0	116 (40)	2364 (52)	
Duration of diabetes (years), mean (SD)	9.2 (10.6)	6.8 (7.6)	<.001
Duration of diabetes (years), n (%)			
<1.0 years	96 (33)	997 (22)	<.001
1.0-9.9 years	89 (30)	2483 (54)	
≥10.0 years	108 (37)	1087 (24)	
Hypertension, n (%)	138 (47)	2324 (51)	.23
Systolic blood pressure (mmHg), mean (SD)	132 (23)	131 (17)	.17
HbA_{1c}[a] (%) (NGSP)[b], mean (SD)	8.8 (2.3)	7.6 (1.3)	<.001
HbA_{1c} (%) (NGSP), n (%)			
<7.0	71 (24)	1444 (32)	<.001
7.0-8.9	101 (34)	2498 (55)	
≥9.0	121 (41)	625 (14)	
eGFR[c] ($mL/min/1.73\ m^2$), mean (SD)	82.3 (31.4)	79.7 (21.0)	.16
eGFR ($mL/min/1.73\ m^2$), n (%)			
<30	17 (6)	47 (1)	<.001
30-59	45 (15)	617 (14)	
≥60	231 (79)	3903 (85)	

[a]HbA_{1c}: glycated hemoglobin.

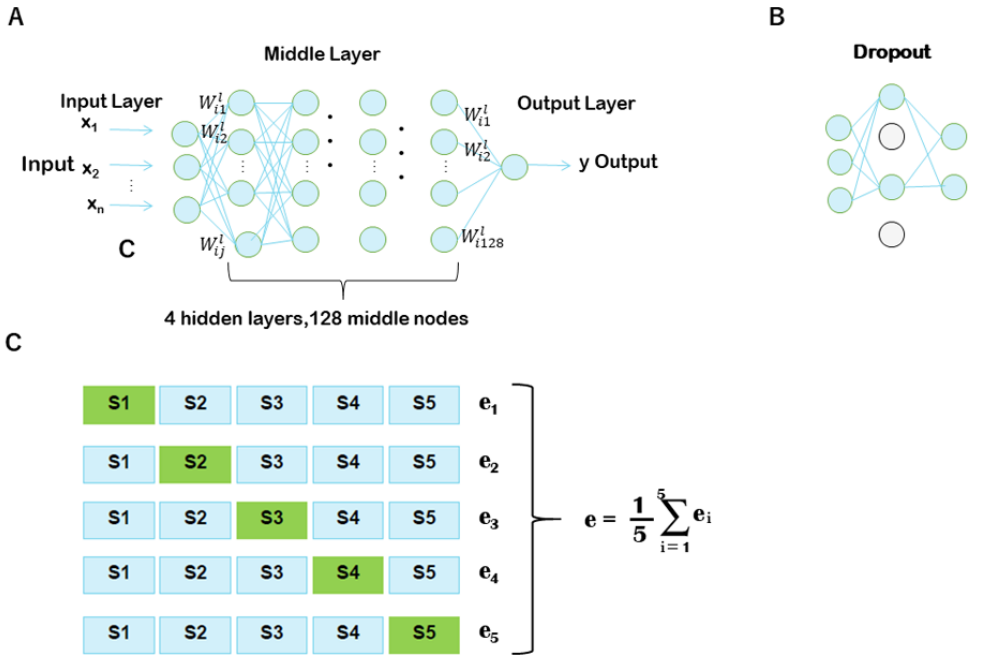
[b]NGSP: National Glycohemoglobin Standardization Program.

[c]eGFR: estimated glomerular filtration rate.

Study 1

We used the full JDDM Study Group data set (N=4860) to evaluate the ability of machine learning models with 5-fold cross-validation analysis for insulin initiation. We divided 4860 prescriptions into 5 groups, maintaining the noninsulin-to-insulin ratio within each group (overall noninsulin-to-insulin ratio of 4567:293). Each training set represented 80% of the data and each test set represented 20% (Figure 1C). We then performed random undersampling, and stratified extraction was adopted. The sampling ratio was verified after being changed to 1:2, 1:4, and 1:8. Specifically, first, using 4860 prescription patterns (ie, using no random undersampling data), the neural network was used to decide on the initial antihyperglycemic medication (insulin or noninsulin initiation). Similarly, using 2576 prescription patterns with a downsampling ratio of 1:2, 1434 prescription patterns with a downsampling ratio of 1:4, and 866 prescription patterns with a downsampling ratio of 1:8, the neural network was used to decide on the initial antihyperglycemic medication. Accuracy, recall, and area under the receiver operating characteristic (ROC) curves (AUCs) were calculated for insulin initiation. Accuracy was defined as the ratio of the sum of the true-positive and true-negative results for all cases. Recall was defined as the ratio of the true-positive cases to the sum of the true-positive and false-negative cases.

Figure 1. (A) Schematic diagram of our neural network models: X=(x1,…, xn) is the input vector and Y=y is the element of the output layer; Wlij is the weight between the ith neuron of the (l)th layer and the jth neuron of the (l-1)th layer . (B) Schematic diagram of dropouts. (C) Schematic diagram of 5-fold cross-validation; S1-S5 indicates data subsets 1 to 5.



Study 2

We compared clinical decisions made by the machine learning approach with those made using logistic regression and by general physicians as to whether to prescribe insulin for patients with type 2 diabetes at the first consultation. We used the full JDDM Study Group data set (N=4860). Forty-three cases that were randomly selected from the 4860 cases to be included in a questionnaire were used for validation data (Multimedia Appendix 1). In random undersampling, stratified extraction was adopted, and the sampling ratios were verified after being changed to 1:2, 1:4, and 1:8. Specifically, first, using 4817 prescription patterns (ie, using no random undersampling data), the neural network and logistic regression were used to decide on the initial antihyperglycemic medication (insulin or noninsulin initiation). Similarly, using 2545 prescription patterns with a downsampling ratio of 1:2, 1409 prescription patterns with a downsampling ratio of 1:4, and 841 prescription patterns with a downsampling ratio of 1:8, the neural network and logistic regression were used to decide on the initial antihyperglycemic medication. In the neural network, each training set represented 80% of the data. We repeated the training 5 times and calculated the average predictive value. The ability of the neural network and logistic analysis to predict insulin initiation in 43 patients was examined according to accuracy, recall, and AUCs.

Study 3

We compared clinical decisions made by the machine learning approach and logistic regression with those made by nonspecialists regarding whether to prescribe insulin for patients with type 2 diabetes at the first consultation, focusing on more definitive cases. In study 3, we evaluated only 7 cases for which the choice of insulin as the initial antidiabetic medication was agreed upon by 8 of the 9 specialists who considered the 43 cases (Multimedia Appendix 2). The ability of a neural network and logistic analysis to predict insulin initiation was evaluated for accuracy.

Questionnaires

This study used a questionnaire to compare the choice of the initiation of each antihyperglycemic drug between general physicians and specialists in clinical settings. We submitted the questionnaire to 50 physicians randomly selected from a list of general physicians (internal medicine physicians) without board certification as diabetologists in Niigata Prefecture; 22 general physicians completed the questionnaire. Nine specialists from university hospitals also completed the same questionnaire. Each physician chose the most suitable antidiabetic drug based on 7 variables (age, sex, BMI, duration of diabetes, HbA_{1c}, hypertension, and eGFR) in 43 cases that were randomly selected from the JDDM Study Group database.

Neural Networks

We used neural networks [19,20] to extract the choice of insulin use by diabetes specialists. A neural network is a mechanism of information processing that emulates the mechanisms of the brain to classify information and identify patterns. Figure 1A is a schematic diagram of our models, where $X=(x_1,\ldots, x_n)$ is the input vector, Y=y is the element of the output layer, and W^l_{ij} is the weight between the *i*th neuron of the (l)th layer and *j*th neuron of the (*l*-1)th layer. Seven explanatory variables (age, sex, BMI, duration of diabetes, HbA_{1c}, hypertension, and eGFR) were used as input nodes (X1-X7), and the output was the predictive value of insulin use by the neural network. Figure 1C is an image of the cross-validation performed in study 1.

For each test, 1 of the 5 subsets was used as the test set and the others were used as training sets. Then, the averages of accuracy, recall, and AUCs across all 5 trials were calculated (study 1). In study 2 and study 3, each training set represented 80% of the data. We repeated the training five times and calculated the average of the predictive value. In this study, because the number of patients who were prescribed insulin was relatively low, we used random undersampling [21,22] to alleviate the imbalance in the data. The numbers used in each random sampling were described above. We used 4 hidden layers, 128 middle nodes, and a rectified linear unit (Relu). Dropouts were also set to suppress overlearning (dropout rate: 0.2 for 1 layer; 0.5 for 2-4 layers) (Figure 1B) [23]. Overlearning was evaluated using learning curve analysis. The number of epochs was validated at 10,000 with the convergence of the difference between accuracy and loss in the learning process. The neural network output was "insulin use," (ie, the predictive value, ranging from 0 to 1 with cutoffs of >0.3, >0.5, and >0.7 for the dichotomous classification of insulin use versus no insulin use in each analysis). The general physicians' choices were compared with the predictions made by machine learning using the neural network for both 43 cases (study 2) and 7 cases (study 3).

Laboratory Data and Definition of Hypertension

HbA_{1c} was converted from the Japanese Diabetes Society's values into the National Glycohemoglobin Standardization Program's equivalent values according to guidelines established by the Japan Diabetes Society [24]. eGFR was determined by an equation modified for the Japanese population as previously described [25]. Hypertension was defined as a systolic blood pressure ≥140 mmHg and/or a diastolic blood pressure ≥90 mmHg, or current use of antihypertensive agents.

Statistical Analysis

Categorical variables were expressed as numerals and percentages and were compared with χ^2 tests. Continuous variables were expressed as mean (SD) and were compared using the Student *t* test for comparisons within each group. Differences in accuracy between general physicians' decisions and the decisions of logistic regression and machine learning were analyzed using the McNemar test. All statistical analyses were performed using SPSS software (version 19.0; IBM Corp) or Python programming. *P* values <.05 were considered statistically significant.

Results

Baseline Characteristics

Table 1 shows participants' baseline characteristics. The number of participants receiving each treatment is shown in Multimedia Appendix 3. With the exceptions of sex, prevalence of hypertension, and systolic blood pressure, there were significant differences between the insulin and noninsulin groups. Participants who were prescribed insulin were younger and had lower BMIs, longer durations of diabetes, and worse glycemic control than those who were not prescribed insulin as their initial medication.

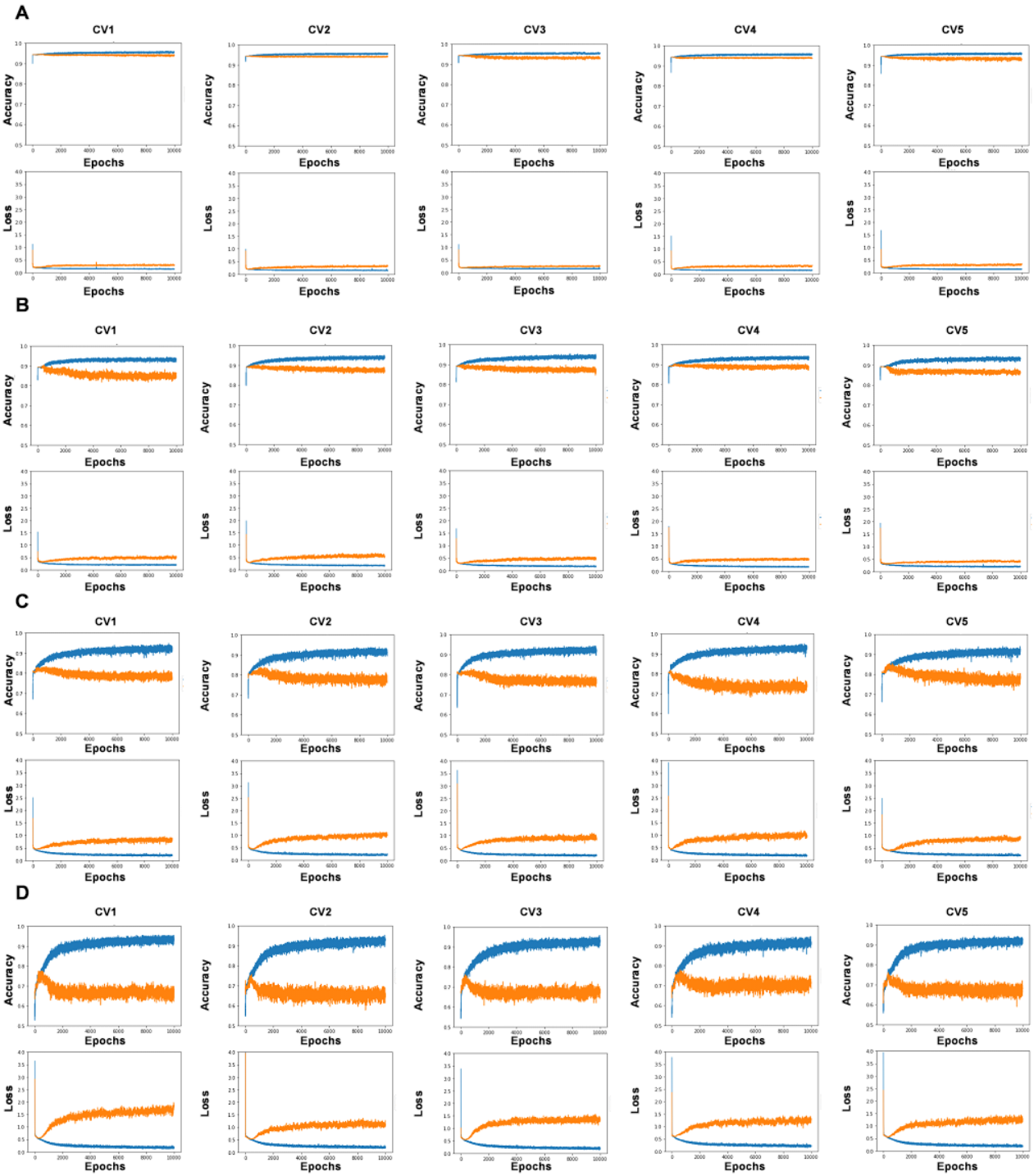
Study 1

Table 2 shows the average accuracy, recall, and AUC of each neural network model using the full JDDM Study Group database (N=4860). Undersampling decreased accuracy but increased recall. AUCs for insulin initiation were approximately 0.6 to 0.7. In learning curve analysis, a tendency of overfitting was observed as the ratio of undersampling increased (Figure 2).

Table 2. Accuracy, recall, and area under the receiver operating characteristic curve (AUC) of each neural network model, with a cutoff of >0.5 for the dichotomous classification.

Cases	Accuracy	Recall	AUC
No undersampling	0.93	0.05	0.61
Sampling ratio 1:2	0.86	0.18	0.63
Sampling rato 1:4	0.78	0.34	0.69
Sampling rato 1:8	0.67	0.45	0.64

Figure 2. Learning curve analysis. (A) No undersampling. (B) Sampling ratio of 1:2. (C) Sampling ratio of 1:4. (D) Sampling ratio of 1:8. The top row shows the association between accuracy and number of epochs, and the bottom row shows the association between cross-entropy loss and number of epochs; the blue and orange lines show the results of training and validation, respectively. CV: cross-validation.



Study 2

Table 3 and Multimedia Appendices 4 and 5 show the accuracy and recall, and Figure 3 shows the ROC curves, of each neural network model and logistic regression in the 43 validation cases. The AUCs of the neural network models for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.67, 0.74, 0.71, and 0.74, respectively, while the AUCs with logistic regression for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.89, 0.89, 0.89, and 0.90, respectively. Accuracy and recall of logistic regression were higher than those of machine learning with a sampling ratio of 1:8. However, the difference in accuracy between the decisions made by logistic regression and machine learning was not statistically significant. Figure 4 shows the learning curve analysis. A tendency of overfitting was observed as the ratio of undersampling increased. The overall accuracy and recall of general physicians were 0.60 and 0.16, respectively. The difference in accuracy between logistic regression and general physicians was statistically significant with a cutoff of >0.5 for the dichotomous classification in the

sampling ratio of 1:8 (P<0.05). We found no statistical significance between machine learning and general physicians.

Table 3. Accuracy and recall of each neural network model and logistic regression with a cutoff of >0.5 for the dichotomous classification.

Models	Accuracy	Recall
Neural network model		
No undersampling	0.60	0.11
Sampling ratio 1:2	0.72	0.37
Sampling ratio 1:4	0.65	0.37
Sampling ratio 1:8	0.70	0.52
Logistic regression		
No undersampling	0.58	0.05
Sampling ratio 1:2	0.65	0.21
Sampling ratio 1:4	0.67	0.26
Sampling ratio 1:8	0.81	0.68

Figure 3. Receiver operating characteristic (ROC) curve of each neural network model and logistic regression for insulin initiation. The areas under the curve of the neural network model (upper row) for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.67, 0.74, 0.71, and 0.74, respectively. For logistic regression (lower row), the areas under the curve for no undersampling, sampling ratio of 1:2, sampling ratio of 1:4, and sampling ratio of 1:8 were 0.89, 0.89, 0.89, and 0.90, respectively.

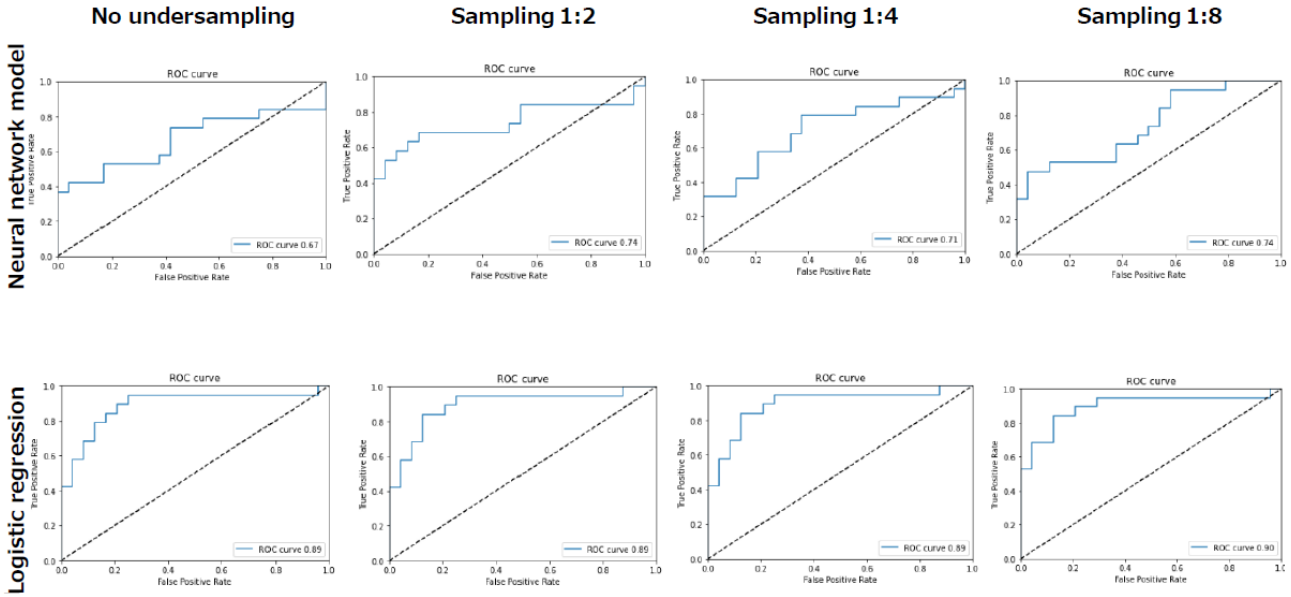
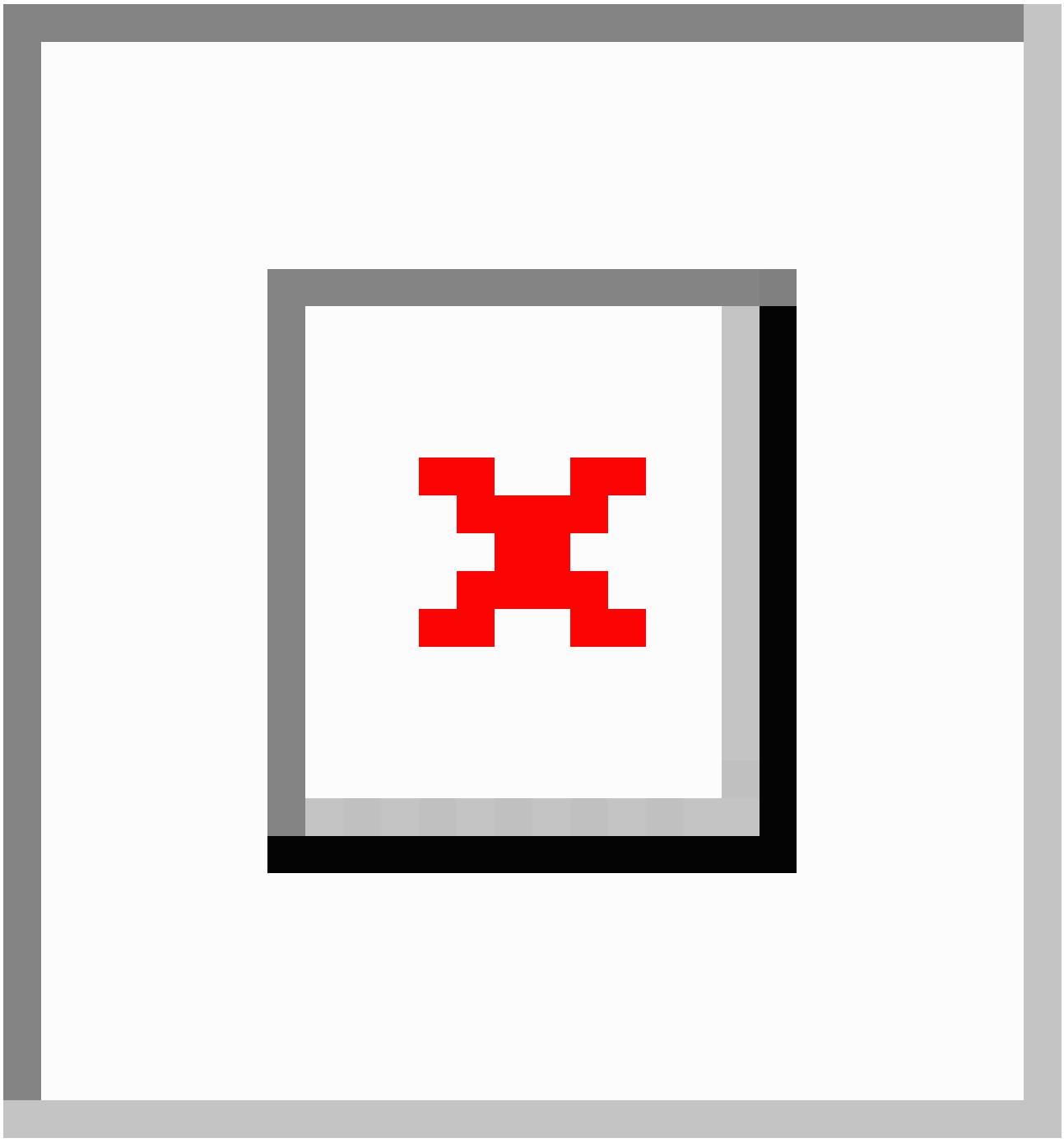


Figure 4. Learning curve analysis. (A) No undersampling. (B) Sampling ratio of 1:2. (C) Sampling ratio of 1:4. (D) Sampling ratio of 1:8. The top row shows the association between accuracy and number of epochs, and the bottom row shows the association between cross-entropy loss and number of epochs; the blue and orange lines show the results of training and validation, respectively.



Study 3

The overall accuracy of insulin initiation by general physicians was 0.51 for the 7 cases for which the choice of insulin as the initial antidiabetic medication was agreed upon by 8 of the 9 specialists. The average predictive values (output) of insulin initiation by machine learning were 0.18, 0.40, 0.50, and 0.82, respectively, for no undersampling and sampling ratios of 1:2, 1:4, and 1:8 (Multimedia Appendix 6). The average predictive values for insulin initiation by logistic regression were 0.38, 0.52, 0.66, and 0.80, respectively, for no undersampling and sampling ratios of 1:2, 1:4, and 1:8 (Multimedia Appendix 6). The accuracies of logistic regression and the machine learning model using 0.5 for the dichotomous classification were 1.00 and 0.86, respectively, with a downsampling ratio of 1:8, which were higher than the accuracy of the general physicians (ie, 0.43) using 50% for the dichotomous classification.

Discussion

Principal Findings

To the best of our knowledge, despite its preliminary stage, this is the first trial to determine whether important clinical decisions, such as the selection of antidiabetic medication, made by a machine learning system could be comparable with

decisions made by diabetes specialists or general physicians. Although we found no superior performance of machine learning over logistic regression, recall in machine learning was relatively similar to that of logistic regression analysis (study 2). In study 3, the accuracy of machine learning with a sampling ratio of 1:8 was higher than that of general physicians. Although further study is needed before machine learning–based decision support systems can be used for insulin initiation in clinical practice, these findings suggest the possibility that machine learning may support such decisions by general physicians.

Barnes et al [26] revealed that models using 7 variables (eg, age, family history of diabetes, BMI, fasting venous glucose level, HbA_{1c}, prior gestational diabetes mellitus, and early diagnosis of gestational diabetes mellitus) could predict required insulin therapy with the addition of medical nutrition therapy in women with gestational diabetes. They showed that the AUC for the prediction of insulin use was 0.71 [26], a value similar to that found in our neural network model. In our study, logistic regression analysis using 7 variables showed that the accuracy and AUC for initial insulin/noninsulin discrimination were consistently higher than with the neural network. A review by Christodoulou et al [27] showed that evidence was lacking to support the claim that clinical prediction models based on machine learning lead to better AUCs than those based on logistic regression. Stylianou et al [28] revealed that an established logistic regression model performed as well as more complex machine learning methods in predictions of mortality from burns. Although recall in machine learning was relatively similar to that of logistic regression analysis in our study, further study is needed before machine learning can be used for decisions on insulin initiation in clinical practice because the neural network model cannot be clearly explained.

In our study, accuracy and recall in logistic regression with a cutoff of >0.3 for the dichotomous classification were higher than with a cutoff of >0.5 although this trend was not observed in the neural network model (Multimedia Appendix 3). Recall was modestly decreased in the neural network model with a cutoff of >0.7 for the dichotomous classification compared with the model with a cutoff of >0.5 for the dichotomous classification. Those findings suggest that with the neural network models, recall might be reduced even with a relatively high cutoff value as a discriminating criterion. However, although insulin initiation is an important clinical decision, recall was relatively low in our neural network model. Therefore, this issue of recall should be resolved before using machine learning–based decision support systems for insulin initiation in clinical practice.

We used random undersampling because the number of patients who were prescribed insulin was relatively low. Also, we attempted to reduce overlearning using dropouts. However, overfitting was still present, especially with the undersampling ratio of 1:8. Thus, no conclusions can be drawn on the usefulness of machine learning as a support system for decisions on insulin initiation until these issues are addressed.

Shortcomings in the accuracy of the prediction of insulin initiation may result from the influence of areas of ambiguity in our study, as there are no absolute standards for insulin initiation. This is in contrast to cancer imaging, for example, where there are consistent gold standards. In summary, the final decision depends on each physician. In fact, predicting insulin initiation through the use of only 7 clinical variables was a limitation mandated by lack of more complete data on our cohort. Predictability of insulin initiation could have been significantly improved if baseline information were available on the symptoms of hyperglycemia, weight loss, metabolic decompensation and ketosis, time course and severity of hyperglycemic symptoms, comorbidities, cardiovascular disease, microvascular complications, dementia, mental disorders, and various results of blood tests, such as C-peptide and glutamic acid decarboxylase antibody. These are key factors in the choice of insulin as initial treatment. Lyons et al [29] showed that initial body weight and peak insulin response were able to predict whether insulin therapy would be required in the subsequent 6 years in symptomatic diabetic patients aged 40 to 60 years with newly diagnosed diabetes. Moreover, doctor and patient values and preferences should be considered in the choice of antihyperglycemic drugs [2]. Since our findings are at a preliminary stage, further studies are needed to produce a tool to support decision making using machine learning in clinical practice, including aspects related to both doctors and patients.

As shown by Case D in Multimedia Appendix 6, machine learning could not predict insulin initiation. The duration of diabetes in Case D was only 0.2 years, which suggests that glucose metabolism worsened in a relatively short period of time. Diabetes specialists choose insulin as the initial therapy to prevent acute exacerbation of glycemic control. Therefore, the findings in Case D indicate that specific cases should be treated with insulin therapy regardless of other clinical variables.

Although we randomly selected a cohort of 43 patients to evaluate the predictability of machine learning, those 43 patients had a lower mean BMI and HbA_{1c} level compared with the entire patient sample (N=4860). The percentages of initial prescriptions of insulin differed between the entire cohort and the 43 randomly selected patients, leading to a discrepancy in the rate of insulin initiation between these two cohorts. Moreover, the insulin-to-noninsulin ratio was not strictly consistent with previous reports [4,19]. In addition, we selected 7 cases as the gold standard based on agreement of 8 of the 9 specialists in study 3. However, the number of validation samples was too small to conclude the usefulness of the ability of machine learning to predict insulin initiation.

The 7 variables in our study were those frequently encountered in clinical settings [4,5]; however, both general physicians and specialists may be unaware that all of these 7 factors could play a role in decisions regarding the use of insulin. Therefore, a simple, automatic, electronic medical system might be useful in addressing this problem. Unfortunately, our findings could not establish the cutoff levels for some variables, such as age, duration of diabetes, BMI, and eGFR, because of the small sample size. Further studies are needed to establish a meaningful decision-making support tool for use in actual consultations with regard to precision medicine.

Study Limitations

Our study has several limitations. First, we randomly selected only 43 samples from the JDDM Study Group database for our questionnaire, as general physicians and specialists are reluctant to respond to long questionnaires. Therefore, the insulin-to-noninsulin ratio was not consistent with that observed by general physicians in clinical settings. Second, although we tried to reduce overlearning using cross-validation and dropouts, overfitting was still present. Thus, our findings should be interpreted with caution. Third, we could not obtain certain information, such as weight loss and hyperglycemic symptoms, that would affect insulin prescriptions because of incomplete data in the CoDiC database. In addition, selection bias was a concern because we included only patients with type 2 diabetes with data available on all 7 variables. In any case, our findings are at a preliminary stage and future studies are needed to produce a decision-making support tool for machine learning in clinical settings that includes those important variables. Fourth, the fact that the study population was exclusively ethnic Japanese may limit wider applicability of the results. Fifth, all of the gold standard cases were males.

Conclusion

Although we found no superior performance of machine learning over logistic regression, machine learning had higher accuracy in the prediction of insulin initiation than general physicians, defined by diabetes specialists' choice of the gold standard. Further study is needed before machine learning–based decision support systems for insulin initiation can be introduced into clinical practice.

Acknowledgments

The authors would like to thank the members of the JDDM Study Group who participated in the study. The authors also thank Kiyoshi Yokoyama from the Graduate School of Science and Technology, Niigata University, for excellent assistance.

This work is supported in part by the Japan Society for the Promotion of Science (19H04028).

We are unable to provide an anonymized data set containing our underlying data used to create the figures and tables because these data are private property of the JDDM Study Group. Making these data available to the general public will result in loss of ownership of the data by the JDDM Study Group.

Authors' Contributions

HS had full access to all of the study data and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study members who contributed significantly to this work are as follows: study concept and design: KF, YM, and MYH; acquisition of data: KF, YM, and MYH; analysis and interpretation of data: KF, YM, MYH, TI, KM, and YH; drafting of the manuscript: KF, YM, MYH, TI, KM, and YH; critical revision of the manuscript for important intellectual content: KF, YM, MYH, MY, TY, HM, SK, and HS; statistical analysis: KF, TI, KM, and YH; and study supervision: HY, YM, MYH, SK, and HS.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Characteristics of study participants in the cohort of 43 patients.
[DOCX File , 16 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Characteristics of study participants for which initial use of insulin was agreed upon by 8 of 9 specialists.
[DOCX File , 15 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Number of patients receiving each hypoglycemia agent.
[DOCX File , 14 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Accuracy and recall of each neural network model and logistic regression.
[DOCX File , 14 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Accuracy and recall of each neural network model and logistic regression.
[DOCX File , 14 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Accuracy and predictive value of each of 7 study participants for insulin initiation by neural network, logistic regression, and general physicians.
[DOCX File , 19 KB-Multimedia Appendix 6]

References

1. American Diabetes Association. Addendum. 9. Pharmacologic Approaches to Glycemic Treatment: Diabetes Care 2020;43(Suppl. 1):S98-S110. Diabetes Care 2020 Aug;43(8):1979. [doi: 10.2337/dc20-ad08a] [Medline: 32503835]
2. Odawara M, Ishii H, Tajima N, Iwamoto Y. Impact of patient attitudes and beliefs to insulin therapy upon initiation, and their attitudinal changes after initiation: the DAWN Japan study. Curr Med Res Opin 2016;32(4):681-686. [doi: 10.1185/03007995.2015.1136605] [Medline: 26743676]
3. Fujihara K, Hanyu O, Heianza Y, Suzuki A, Yamada T, Yokoyama H, et al. Comparison of clinical characteristics in patients with type 2 diabetes among whom different antihyperglycemic agents were prescribed as monotherapy or combination therapy by diabetes specialists. J Diabetes Investig 2016 Mar;7(2):260-269 [FREE Full text] [doi: 10.1111/jdi.12387] [Medline: 27042280]
4. Fujihara K, Igarashi R, Matsunaga S, Matsubayashi Y, Yamada T, Yokoyama H, et al. Comparison of baseline characteristics and clinical course in Japanese patients with type 2 diabetes among whom different types of oral hypoglycemic agents were chosen by diabetes specialists as initial monotherapy (JDDM 42). Medicine 2017;96(7):e6122. [doi: 10.1097/md.0000000000006122]
5. Grant RW, Wexler DJ, Watson AJ, Lester WT, Cagliero E, Campbell EG, et al. How doctors choose medications to treat type 2 diabetes: a national survey of specialists and academic generalists. Diabetes Care 2007 Jun;30(6):1448-1453 [FREE Full text] [doi: 10.2337/dc06-2499] [Medline: 17337497]
6. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA 2017 Aug 08;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
7. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
8. Abhari S, Niakan Kalhori SR, Ebrahimi M, Hasannejadasl H, Garavand A. Artificial Intelligence Applications in Type 2 Diabetes Mellitus Care: Focus on Machine Learning Methods. Healthc Inform Res 2019 Oct;25(4):248-261 [FREE Full text] [doi: 10.4258/hir.2019.25.4.248] [Medline: 31777668]
9. Contreras I, Vehi J. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. J Med Internet Res 2018 May 30;20(5):e10775 [FREE Full text] [doi: 10.2196/10775] [Medline: 29848472]
10. Verbraak FD, Abramoff MD, Bausch GC, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic Accuracy of a Device for the Automated Detection of Diabetic Retinopathy in a Primary Care Setting. Diabetes Care 2019 Apr;42(4):651-656. [doi: 10.2337/dc18-0148] [Medline: 30765436]
11. Haenssle H, Fink C, Rosenberger A, Uhlmann L. Reply to the letter to the editor 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists' by H. A. Haenssle et al. Ann Oncol 2019 May 01;30(5):854-857. [doi: 10.1093/annonc/mdz015] [Medline: 30689691]
12. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
13. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Prog Retin Eye Res 2018 Nov;67:1-29 [FREE Full text] [doi: 10.1016/j.preteyeres.2018.07.004] [Medline: 30076935]
14. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
15. Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. Cancers (Basel) 2019 Aug 23;11(9) [FREE Full text] [doi: 10.3390/cancers11091235] [Medline: 31450799]
16. Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. Br J Radiol 2018 Sep;91(1089):20170545 [FREE Full text] [doi: 10.1259/bjr.20170545] [Medline: 29565644]
17. Kobayashi M, Yamazaki K, Hirao K, Oishi M, Kanatsuka A, Yamauchi M, Japan Diabetes Clinical Data Management Study Group. The status of diabetes control and antidiabetic drug therapy in Japan--a cross-sectional survey of 17,000 patients with diabetes mellitus (JDDM 1). Diabetes Res Clin Pract 2006 Aug;73(2):198-204. [doi: 10.1016/j.diabres.2006.01.013] [Medline: 16621117]
18. Oishi M, Yamazaki K, Okuguchi F, Sugimoto H, Kanatsuka A, Kashiwagi A, Japan Diabetes Clinical Data Management Study Group. Changes in oral antidiabetic prescriptions and improved glycemic control during the years 2002-2011 in Japan (JDDM32). J Diabetes Investig 2014 Sep;5(5):581-587 [FREE Full text] [doi: 10.1111/jdi.12183] [Medline: 25411627]

19. Jayanta KB, Debnath B, Tai-hoon K. Use of Artificial Neural Network in Pattern Recognition. International Journal of Software Engineering and Its Applications 2010;4:23-34.
20. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015 Jan;61:85-117. [doi: 10.1016/j.neunet.2014.09.003] [Medline: 25462637]
21. Holte R. C4. 2003 Presented at: 5, class imbalance,cost sensitivity: why under-sampling beats over-sampling. . In Workshop on learning from imbalanced datasets II. ; 11; 2003; Washington DC p. 1-8.
22. Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies. 2000:AAAI Technical Report WS-2000:00.
23. Srivastava N, Hinton GA. K, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014;15:A.
24. Kashiwagi A, Kasuga M, Araki E, Oka Y, Hanafusa T, Ito H. International clinical harmonization of glycated hemoglobin in Japan: From Japan Diabetes Society to National Glycohemoglobin Standardization Program values. Diabetology International 2012;3:8-10. [doi: 10.1007/s13340-012-0069-8]
25. Matsuo S, Imai E, Horio M, Yasuda Y, Tomita K, Nitta K. Revised equations for estimated GFR from serum creatinine in Japan. Am J Kidney Dis 2009;53:982-992. [doi: 10.1053/j.ajkd.2008.12.034]
26. Barnes R, Wong T, Ross G, Jalaludin B, Wong V, Smart C. A novel validated model for the prediction of insulin therapy initiation and adverse perinatal outcomes in women with gestational diabetes mellitus. Diabetologia 2016;59:2331-2338. [doi: 10.1007/s00125-016-4047-8]
27. Christodoulou E, Ma J, Collins G, Steyerberg E, Verbakel J, Van CB. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004]
28. Stylianou N, Akbarov A, Kontopantelis E, Buchan I, Dunn K. Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. Burns : journal of the International Society for Burn Injuries 2015;41:925-934. [doi: 10.1016/j.burns.2015.03.016]
29. Lyons T, Kennedy L, Atkinson A, Buchanan K, Hadden D, Weaver J. Predicting the need for insulin therapy in late onset (40-69 years) diabetes mellitus. Diabet Med 1984;1:105-107. [doi: 10.1111/j.1464-5491.1984.tb01938.x]

Abbreviations

AUC: area under the receiver operating characteristic curve
eGFR: estimated glomerular filtration rate
HbA1c: glycated hemoglobin
JDDM: Japan Diabetes Clinical Data Management
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 05.07.20; peer-reviewed by A Masino, H Zhang; comments to author 07.10.20; revised version received 23.11.20; accepted 12.12.20; published 27.01.21

Please cite as:
Fujihara K, Matsubayashi Y, Harada Yamada M, Yamamoto M, Iizuka T, Miyamura K, Hasegawa Y, Maegawa H, Kodama S, Yamazaki T, Sone H
Machine Learning Approach to Decision Making for Insulin Initiation in Japanese Patients With Type 2 Diabetes (JDDM 58): Model Development and Validation Study
JMIR Med Inform 2021;9(1):e22148
URL: http://medinform.jmir.org/2021/1/e22148/
doi: 10.2196/22148
PMID: 33502325