

Original Paper

Treatment Discontinuation Prediction in Patients With Diabetes Using a Ranking Model: Machine Learning Model Development

Hisashi Kurasawa^{1*}, PhD; Kayo Waki^{2*}, MPH, MD, PhD; Akihiro Chiba^{1,3}, PhD; Tomohisa Seki², MD, PhD; Katsuyoshi Hayashi¹, PhD; Akinori Fujino¹, PhD; Tsuneyuki Haga^{1,4}, PhD; Takashi Noguchi⁵, MD, PhD; Kazuhiko Ohe², MD, PhD

¹Nippon Telegraph and Telephone Corporation, Tokyo, Japan

²Department of Healthcare Information Management, The University of Tokyo Hospital, Tokyo, Japan

³NTT DOCOMO, INC, Tokyo, Japan

⁴NTT-AT IPS Corporation, Kanagawa, Japan

⁵National Center for Child Health and Development, Tokyo, Japan

*these authors contributed equally

Corresponding Author:

Kayo Waki, MPH, MD, PhD

Department of Healthcare Information Management

The University of Tokyo Hospital

7-3-1 Hongo, Bunkyo-ku

Tokyo, 113-8655

Japan

Phone: 81 3 5800 9077

Email: kwaki-tyk@m.u-tokyo.ac.jp

Abstract

Background: Treatment discontinuation (TD) is one of the major prognostic issues in diabetes care, and several models have been proposed to predict a missed appointment that may lead to TD in patients with diabetes by using binary classification models for the early detection of TD and for providing intervention support for patients. However, as binary classification models output the probability of a missed appointment occurring within a predetermined period, they are limited in their ability to estimate the magnitude of TD risk in patients with inconsistent intervals between appointments, making it difficult to prioritize patients for whom intervention support should be provided.

Objective: This study aimed to develop a machine-learned prediction model that can output a TD risk score defined by the length of time until TD and prioritize patients for intervention according to their TD risk.

Methods: This model included patients with diagnostic codes indicative of diabetes at the University of Tokyo Hospital between September 3, 2012, and May 17, 2014. The model was internally validated with patients from the same hospital from May 18, 2014, to January 29, 2016. The data used in this study included 7551 patients who visited the hospital after January 1, 2004, and had diagnostic codes indicative of diabetes. In particular, data that were recorded in the electronic medical records between September 3, 2012, and January 29, 2016, were used. The main outcome was the TD of a patient, which was defined as missing a scheduled clinical appointment and having no hospital visits within 3 times the average number of days between the visits of the patient and within 60 days. The TD risk score was calculated by using the parameters derived from the machine-learned ranking model. The prediction capacity was evaluated by using test data with the C-index for the performance of ranking patients, area under the receiver operating characteristic curve, and area under the precision-recall curve for discrimination, in addition to a calibration plot.

Results: The means (95% confidence limits) of the C-index, area under the receiver operating characteristic curve, and area under the precision-recall curve for the TD risk score were 0.749 (0.655, 0.823), 0.758 (0.649, 0.857), and 0.713 (0.554, 0.841), respectively. The observed and predicted probabilities were correlated with the calibration plots.

Conclusions: A TD risk score was developed for patients with diabetes by combining a machine-learned method with electronic medical records. The score calculation can be integrated into medical records to identify patients at high risk of TD, which would be useful in supporting diabetes care and preventing TD.

KEYWORDS

machine learning; machine-learned ranking model; treatment discontinuation; diabetes; prediction; electronic health record; EHR; big data; ranking; algorithm

Introduction

Background

Diabetes is a chronic disease requiring both self-management and long-term management. Poor glycemic control increases the risk of complications, including cardiovascular and cerebrovascular diseases as well as macrovascular and microvascular diseases, such as nephropathy, retinopathy, and neuropathy [1-4]. To prevent the progression of these complications, adherence to dietary, exercise, and medication regimens is necessary [5]. Nonadherence has been shown to increase the risk of morbidity [4] and all-cause mortality [6].

Treatment discontinuation (TD), defined as dropping out of regular medical care, is likely to result in the worsening of glycemic control and progression of complications [3,4]. TD rates in patients with diabetes are rather high, ranging from 4% to 19% in the United Kingdom [3,4], 12% to 50% in the United States [7,8], and 13.5% to 56.9% in Japan [9,10]. Furthermore, patients who have previously discontinued treatment have been shown to have a 3-fold higher risk of repeated TD than those who have never done so [11].

Prior Work

Preventing TD is crucial in the management of diabetes, and several studies have statistically analyzed the factors associated with TD [6-8,12]. Previously identified factors include younger age [6,13], smoking [6,14], poor glycemic control [6,13,15,16], high blood pressure [13], obesity [9], medications [12,16], employment status [8,17], region [18], transportation barriers [7,19,20], clinical appointments [20], and complications [21]. The most commonly used statistical hypothesis tests are *t* test and chi-square test. However, a review [22] pointed out a variety of multilevel factors in association with TD with inconsistent findings. It has remained difficult for clinicians to carefully discern each patient's risk of TD.

Machine learning (ML) may be useful for predicting each patient's risk of TD by taking into account a wide variety of factors. Statistics focus on *explaining outcomes with data*, whereas ML focuses on *predicting outcomes with data* [23]. Although ML cannot identify consistent factors, it can inform clinicians about who is a high-risk patient for TD. It could help clinicians shift their time spent on identifying high-risk patients to encouraging them to continue treatment. According to a systematic review by Carreras-García et al [24], most studies designed their model as a binary classification problem [25] that classified scheduled appointments based on whether they were kept or missed. Furthermore, the most commonly used model was logistic regression, and the most frequently used metric was the area under the receiver operating characteristic curve (AUROC). However, as a binary classification outputs the probability of a missed appointment (MA) occurring after a predetermined period, it is limited in its ability to estimate the

magnitude of TD risk in patients with inconsistent intervals between appointments. Even if a patient missed an appointment, if the frequency of visits was maintained such that their condition did not worsen thereafter, the TD risk of the patient would be low. An MA is a necessary but not sufficient condition for TD.

Goal of This Study

In this study, we aimed to develop a novel method of calculating TD risk via ML. We designed a prediction model of TD as a ranking problem with imbalanced data to compare patients by length of time until TD. The ranking problem [26] is an application of survival time analysis [27]. Cox regression [28] is generally used in statistical analysis, whereas the ranking model is used in ML [29-31]. Cox regression is a model of the hazard function in which the effects of the explanatory variables on outcomes are predetermined, requiring an assumption that they remain constant over time [28]. In contrast, the ranking model does not require this assumption and makes flexible use of the variables. Furthermore, because there was a concern that the learning model would have a heavier bias toward TD cases than treatment continuation (TC) cases, the sampling was devised on the basis of the findings of the imbalanced data.

The contributions of this work are as follows:

1. This study designed a prediction model of TD as a ranking problem with imbalanced data, which allows for a comparison of patients' risk of TD with the time remaining before TD. This is the first study to use a machine-learned ranking model to predict TD.
2. The mean (95% confidence limits) of the C-index for the TD risk score obtained with the model was 0.749 (0.655, 0.823). This was higher than 0.662 (0.574, 0.748), which was obtained with the Cox regression model; the results for the AUROC and area under the precision-recall curve (AUPRC) were similar.

Methods

Ethics Approval

This study was approved by the research ethics committees of the Graduate School of Medicine and Faculty of Medicine at the University of Tokyo (approval number: 10705) and was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained, and an opportunity to opt out of participation was provided.

Study Population

All data were collected from electronic health records (EHRs) at the University of Tokyo Hospital, which included 7551 patients who visited the hospital after January 1, 2004, and had diagnostic codes indicative of diabetes. Characteristics of patient in the training and test data are shown in [Table 1](#).

Table 1. Characteristics of patients in the training and test data.

Characteristics	Training data (n=6509)		Test data (n=1042)	
	TD ^a (n=204, 3.13%)	TC ^b (n=6305, 96.86%)	TD (n=38, 3.65%)	TC (n=1004, 96.35%)
Number of appointments, mean (SD)	4.8 (3.3)	10.4 (5.0)	3.1 (2.6)	5.8 (4.1)
Number of missed appointments, mean (SD)	1.6 (1.2)	1.6 (1.2)	1.2 (0.5)	1.3 (0.7)
Age (years), mean (SD)	62.6 (15.9)	66.0 (12.6)	59.9 (15.0)	61.1 (14.1)
<20, n (%)	0 (0)	3 (0.05)	0 (0)	1 (0.10)
20-30, n (%)	5 (2.50)	45 (0.71)	1 (3)	25 (2.49)
30-40, n (%)	14 (6.90)	204 (3.24)	4 (11)	63 (6.27)
40-50, n (%)	28 (13.70)	452 (7.17)	6 (16)	117 (11.65)
50-60, n (%)	31 (15.20)	883 (14)	6 (16)	188 (18.73)
60-70, n (%)	47 (23)	1950 (30.93)	8 (21)	310 (30.88)
≥70, n (%)	79 (38.70)	2768 (43.90)	13 (34)	300 (29.88)
Sex, n (%)				
Male	127 (63.30)	3777 (59.90)	25 (66)	594 (59.16)
Female	77 (37.70)	2528 (40.10)	13 (34)	410 (40.84)
Hospital visit interval in days, mean (SD)	65.9 (33.1)	57.3 (23.9)	56.2 (65.5)	49.0 (21.0)
<30, n (%)	4 (2)	283 (4.49)	7 (18)	127 (12.65)
30-60, n (%)	72 (35.30)	3237 (51.34)	15 (39)	511 (50.90)
60-90, n (%)	66 (32.30)	2140 (33.94)	3 (8)	177 (17.63)
≥90, n (%)	26 (12.80)	415 (6.58)	2 (5)	39 (3.88)
First visit, n (%)	36 (17.70)	230 (3.65)	11 (29)	150 (14.94)
HbA_{1c}^c (NGSP^d), %, mean (SD)	7.1 (1.2)	7.0 (1.0)	7.0 (1.1)	7.0 (1.1)
<6, n (%)	31 (15.20)	770 (12.21)	6 (16)	118 (11.75)
6-7, n (%)	64 (31.40)	2281 (36.18)	12 (32)	382 (38.05)
7-8, n (%)	48 (23.50)	1788 (28.36)	9 (24)	285 (28.39)
≥8, n (%)	33 (16.20)	632 (10.02)	4 (11)	148 (14.74)
Missing value, n (%)	28 (13.70)	834 (13.23)	7 (18)	71 (7.07)
TG^e, mg/dL, mean (SD)	182.2 (167.4)	143.5 (96.5)	199.0 (239.1)	160.5 (120.9)
<30, n (%)	0 (0)	4 (0.06)	0 (0)	0 (0)
30-150, n (%)	91 (44.60)	3601 (57.11)	15 (39)	550 (54.78)
150-300, n (%)	65 (31.90)	1631 (25.87)	10 (26)	291 (28.98)
300-750, n (%)	16 (7.80)	213 (3.38)	3 (8)	72 (7.17)
≥750, n (%)	3 (1.50)	11 (0.17)	1 (3)	6 (0.60)
Missing value, n (%)	29 (14.20)	845 (13.40)	9 (24)	85 (8.47)
HDL^f, mg/dL, mean (SD)	58.6 (15)	60.6 (16.9)	54.4 (20.3)	56.6 (16.8)
<20, n (%)	0 (0)	2 (0.03)	0 (0)	0 (0)
20 to <40, n (%)	15 (7.40)	387 (6.14)	8 (21)	130 (12.95)
40 to <100, n (%)	159 (77.90)	4882 (77.43)	20 (52)	759 (75.60)
≥100, n (%)	3 (1.50)	126 (2)	1 (3)	15 (1.49)
Missing value, n (%)	27 (13.20)	908 (14.40)	9 (24)	100 (9.96)

Characteristics	Training data (n=6509)		Test data (n=1042)	
	TD ^a (n=204, 3.13%)	TC ^b (n=6305, 96.86%)	TD (n=38, 3.65%)	TC (n=1004, 96.35%)
LDL^g, mg/dL, mean (SD)	121.6 (31.3)	111.6 (26.8)	119.9 (33.7)	113.0 (35.0)
<60, n (%)	2 (1)	107 (1.70)	1 (3)	26 (2.59)
60-120, n (%)	64 (31.40)	2700 (42.82)	7 (18)	338 (33.67)
120-140, n (%)	36 (17.70)	988 (15.67)	2 (5)	125 (12.45)
≥140, n (%)	32 (15.70)	532 (8.44)	5 (13)	120 (11.95)
Missing value, n (%)	70 (34.30)	1978 (31.37)	23 (61)	395 (39.34)
TCho^h, mg/dL, mean (SD)	201.6 (44.5)	189.5 (32.8)	193.3 (36.6)	192.9 (43.4)
<130, n (%)	2 (1)	152 (2.41)	1 (3)	50 (4.98)
130-220, n (%)	111 (54.40)	4202 (66.65)	20 (53)	650 (64.74)
220-240, n (%)	23 (11.30)	516 (8.18)	6 (16)	97 (9.66)
240-280, n (%)	15 (7.40)	246 (3.90)	1 (3)	77 (7.67)
≥280, n (%)	5 (2.50)	43 (0.68)	0 (0)	29 (2.89)
Missing value, n (%)	48 (23.50)	1146 (18.18)	10 (26)	101 (10.06)

^aTD: treatment discontinuation.

^bTC: treatment continuation.

^cHbA_{1c}: hemoglobin A_{1c}.

^dNGSP: National Glycohemoglobin Standardization Program.

^eTG: triglyceride.

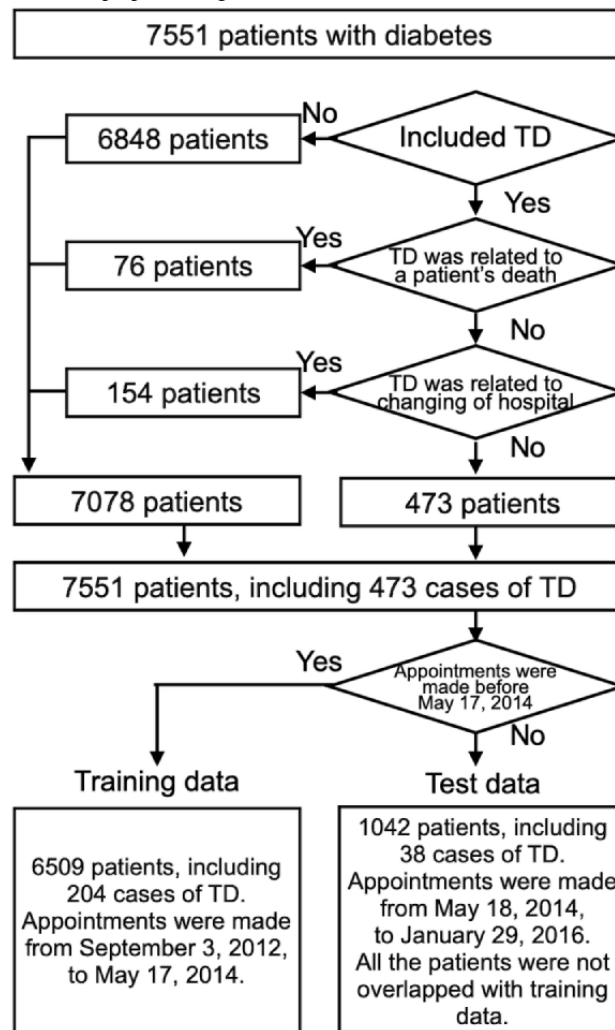
^fHDL: high-density lipoprotein.

^gLDL: low-density lipoprotein.

^hTCho: total choline.

The data were recorded in the EHRs between September 3, 2012, and January 29, 2016. As illustrated in [Figure 1](#), based on the calendar date, two-thirds of the data (days: 828/1243, 66.6%) were used for training (between September 3, 2012, and May 17, 2014) and the remaining one-third (days: 415/1243, 33.4%) was used for testing (between May 18, 2014, and

January 29, 2016). The records used for training were not used for testing to ensure that the same patients were not included in both groups. A total of 6509 patients (204 cases of TD) were included in the training group, and 1042 patients (38 cases of TD) were included in the testing group.

Figure 1. Illustration of patient selection and data preprocessing. TD: treatment discontinuation.

Definition of TD

The TD of a patient was defined as missing a scheduled clinical appointment and having no hospital visits within 3 times the average number of days between the visits of the patient and within 60 days. Each patient's average number of days between visits was calculated from the last 3 visit days. In other words, if 3 times the average number of days between visits was greater than 60 days, then 60 days was used as the threshold. Otherwise, 3 times the average number of days between visits was used as the threshold.

Other studies have defined TD as the lack of hospital visits over a particular threshold of time (between 1 day and 6 months) [6-8,12-21]. When the threshold was set at 60 days, 336 cases of TD were detected in the training data and 65 cases of TD were detected in the test data, but there was a trend that patients with longer visit intervals were more likely to be judged as TD cases. It is not easy to set appropriate thresholds for outpatients whose hospital visits are at inconsistent intervals. Next, when the threshold was set to 3 times the average number of days between visits, 218 cases of TD were detected in the training data and 54 cases of TD were detected in the test data, but

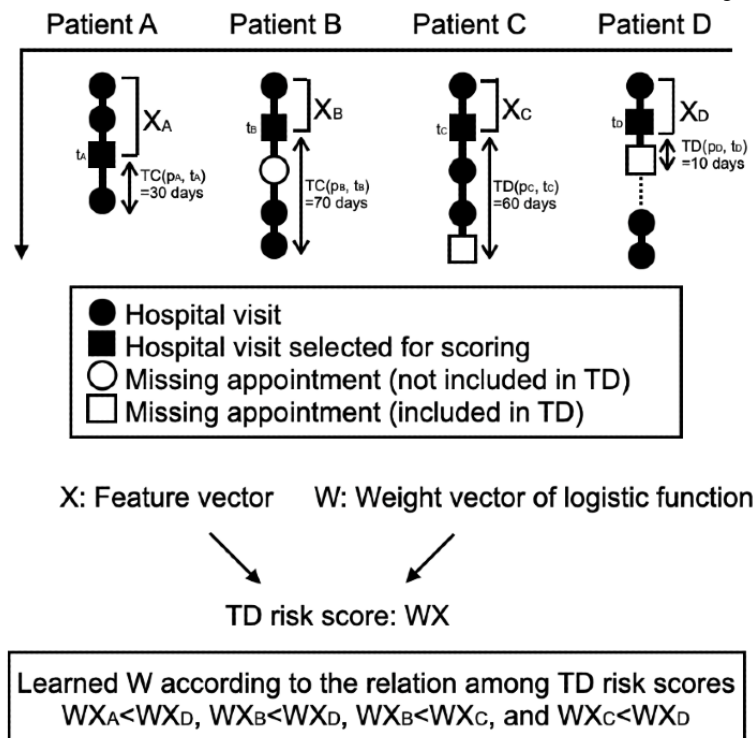
patients with shorter visit intervals tended to be more likely to be judged as TD cases or judged as having a risk of TD. Therefore, we included both conditions in the definition.

To ensure accurate TD detection, a physician, one of the coauthors, verified that the above definition was met and excluded cases of patient death or changes in care setting.

Length of Treatment Until Discontinuation

Length of treatment was measured in 2 ways. First, TD (p_m, t_m) was defined as the number of days from the date t_m to the missed scheduled clinical appointment associated with TD for the patient p_m who had TD (or possible TD). In the second way, TC (p_n, t_n) was defined as the number of days from the date t_n to the most recently recorded visit for the patient p_n who had no TD.

For example, as shown in Figure 2, in the case of patient A, there were 30 days from t_A to the most recently recorded visit, so TC (p_A, t_A) was set to 30 days. In the case of patient C, there were 60 days from t_C to the missed scheduled clinical appointment associated with TD, so TD (p_C, t_C) was set to 60 days.

Figure 2. Examples of the value of the treatment discontinuation (TD) risk. TC: treatment continuation; W: weight vector; X: feature vector.

Class Design

The classification $y_{m,n}$ was based on the difference between a pair of treatment lengths. Here, $y_{m,n}=+1$ for the pair of TD (p_m, t_m) for the patient p_m and the date t_m and TD (p_n, t_n) for the patient p_n and the date t_n if TD (p_m, t_m) is shorter than TD (p_n, t_n) and the pair of TD (p_m, t_m) and TC (p_n, t_n) if TD (p_m, t_m) is shorter than TC (p_n, t_n). $y_{m,n}=-1$ for the pair of TD (p_m, t_m) and TD (p_n, t_n) if TD (p_m, t_m) is longer than TD (p_n, t_n) and for the pair of TC (p_m, t_m) and TD (p_n, t_n) if TC (p_m, t_m) is longer than TD (p_n, t_n).

The classification was performed only when the patients had different times until TD, or when one patient had TD and the other had TC, where TC (p_n, t_n) was longer than TD (p_m, t_m). The classification was not performed on other occasions because the difference in time until TD between the 2 patients could not be compared. For the examples shown in Figure 2, the classes of the pair of TC (p_A, t_A) and TD (p_D, t_D) and that of TC (p_B, t_B) and TD (p_D, t_D), TC (p_B, t_B) and TD (p_C, t_C), and TD (p_C, t_C) and TD (p_D, t_D) were all set to -1 .

Feature Design

To ensure that the factors related to TD were included, we designed a feature vector x_n for patient p_n at time t_n , representing the clinical conditions beginning with the initial visit and lasting until just before t_n . In total, 149,699 features, 51,778 qualitative

features and 97,921 quantitative features, were used. Table 2 describes the features used for the prediction.

We designed the features using 3 classes of representation. The first included detailed demographic and clinical conditions (sex, age, previously consulted medical departments, diagnosed diseases, and prescribed medications). These had numerous features, most of which had a 0 value, leading to a very sparse representation.

The second class included changes occurring during the treatment of a patient to identify the risk of TD at each hospital visit. For example, we used the accumulated number of hospital visits, length of prescription time, number of medications prescribed, laboratory results, day of the week an appointment was scheduled, the interval between the date on which a clinical appointment was made and the scheduled appointment date, and the weather conditions on the appointment day. Detailed histories of hospital visits were included because features related to when and how appointments were made influenced the accuracy of the predicted MAs in our previous work [25].

The third class included data from public databases beyond the EHR. For instance, to represent the distance from a patient's home to the hospital, we used a geographic information system and measured the distance and travel time. We also used information regarding patient occupations. The observed values of each quantitative variable, for example, blood test results, were linearly transformed (normalized) to make the variance of each variable equal to 1. The transformed variable was then assigned to the vector.

Table 2. Description of explanatory variables used for prediction.

Primary and secondary categories	Qualitative variables (n=51,778), n (%)	Quantitative variables (n=97,921), n (%)	Characteristic feature (reference)
Attribute			
Sex and age	4 (0.01)	5 (0.01)	Sex and age
Address	492 (0.95)	492 (0.50)	Distance and time duration from the house to the hospital by public transport (geographic information system)
Insurance	67 (0.13)	3 (0)	Business-type category (health insurance societies of companies)
Consultation			
Medical department, outpatient, and inpatient	267 (0.52)	514 (0.52)	Previously and recently consulted medical departments
Subject	8021 (15.49)	13,108 (13.39)	Subject categories of consultation assigned by each medical department
Time	33 (0.06)	105 (0.11)	Late arrival for an appointment
Appointment (intervals and changes)	74 (0.14)	197 (0.20)	Interval between the date on which a clinical appointment was made and scheduled appointment date
Medicine			
Directions of each medicine	10,346 (19.98)	17,678 (18.05)	How many times a day medication is taken
Doses of each medicine	4570 (8.83)	33,403 (34.11)	Total amount of medication per day
Component	2332 (4.50)	5082 (5.19)	Component (medicine code defined by the Ministry of Health, Labor and Welfare)
Medical department, outpatient, and inpatient	324 (0.63)	678 (0.69)	Medication for outpatient to the department of Diabetes and Metabolic Diseases
Disease (recovered from and under treatment)	21,977 (42.44)	22,012 (22.48)	Disease category under care and recovered (ICD-10 ^a)
Laboratory tests			
Medical department, outpatient, and inpatient	170 (0.33)	357 (0.36)	HbA _{1c} ^b , HDL-C ^c , LDL-C ^d , TG ^e , TChol ^f , etc
Order, exam and intervals	219 (0.42)	462 (0.47)	Interval between tests
Results	297 (0.57)	658 (0.67)	Categorized result according to the criteria (Diabetes Medical Guideline)
Physiological tests (order, exam, and intervals)	2237 (4.32)	2801 (2.86)	Interval between tests
Surgery (procedure)	336 (0.65)	338 (0.35)	Procedure name
Nutritional guidance (medical department, outpatient, and inpatient)	12 (0.05)	28 (0.03)	Guidance for inpatient to the department of Diabetes and Metabolic Diseases

^aICD-10: International Classification of Diseases, Tenth Revision.

^bHbA_{1c}: hemoglobin A_{1c}.

^cHDL-C: high-density lipoprotein.

^dLDL-C: low-density lipoprotein.

^eTG: triglycerides.

^fTChol: total choline.

All the features were generated by processing variables obtained from the EHRs. The category with the highest number of variables was medicine. Raw categorical variables such as medicine name, component, units, inpatient and outpatient category, and department that prescribed the medicine were extracted. Raw numerical variables such as amount, dosage, and number of days or times were extracted. In addition, new

numerical variables were generated by combining categorical and numeric variables such as pairs of medicine name and amount, pairs of medicine name and dosage, and pairs of medicine name and number of days or times. New categorical variables such as pairs of medicine name and inpatient and outpatient category and pairs of medicine name and department were also generated. The category with the second highest

number of features was disease. Raw categorical variables such as disease name; disease category defined by International Classification of Diseases, Tenth Revision; treatment status (under treatment and recovering); and disease type (primary disease and secondary disease) were extracted. In addition, new categorical variables such as pairs of disease name and treatment status and pairs of disease name and disease type were generated. New numerical variables were also generated by counting the number of diseases that were under treatment and recovered for each disease category. The variables of the other categories were as follows. From the attribute category, categorical variables such as sex, names of regions and cities, insurance categories, and business-type categories were extracted. Numerical variables such as age and copayment rates were extracted. Distance and travel time were generated as new numerical variables using geographic information system from region and city names, as described in the third representation class. From the consultation category, categorical variables such as department, inpatient and outpatient category, and subject name of the reservation slot were extracted. Numerical variables such as time of arrival, appointment, clinic start, and clinic end were extracted. These time intervals were generated as new numerical variables. From the appointment category, categorical variables such as department and appointment status (new, change, and cancellation) were extracted. Numerical variables such as time of registration and reservation were extracted. The new numerical variables were generated, as described in the second representation class. From the laboratory and physiological tests categories, categorical variables such as test name, department, and inpatient and outpatient category were extracted. Numerical variables such as test values were extracted. From the surgery category, categorical variables such as operative name were extracted. From the nutritional guidance category, categorical variables such as department and inpatient and outpatient categories were extracted.

Most features were generated using the following 3-step procedure. First, raw variables were extracted from each category, tied to their recorded times, and classified into categorical variables (eg, names of diagnosed diseases) and numeric variables (eg, number of medicines prescribed). Second, Categorical variables were further classified into raw categorical variables and frequency-transformed categorical variables. Third, the combinations of the raw categorical variables and the statistics of the frequency-transformed categorical variables were computed with varying window sizes to generate qualitative features and quantitative features, respectively. Numeric variables were transformed to linear and logarithmic scales, and their statistics were computed with varying window sizes to generate quantitative features. 4 statistics were used for feature generation: minimum, maximum, mean, and SD. To relate the most recent trends in circumstances to the TD risk score, periods of 3 months, 6 months, and 1 year before the target time were used as window sizes. A categorical variable was also added to indicate missing data if a feature was present for a shorter time than the window size.

For example, from the attribute category, the features sex, age, address, and insurance were extracted to express demographic conditions. The features of sex consisted of 1 qualitative variable

representing male or female, 3 quantitative variables representing its frequencies with the 3 window sizes, and 3 qualitative variables representing their missing values. The frequencies of the sex variable itself have no meaning, but because it is a variable that is always listed in each EHR, it was used to represent the number of EHRs in the window size. The features of age consisted of 2 quantitative variables of linear and logarithmic scales. The features of address consisted of 48 quantitative variables of the 4 statistics of the 2 scales of the distance and travel time from a patient's home to the hospital with 3 window sizes, 48 qualitative variables representing their missing values, and 444 quantitative and qualitative variables representing the names of regions and cities and their frequencies. The features of insurance consisted of 67 qualitative variables representing insurance categories and business-type categories and 3 quantitative variables representing copayment rates.

Model Design

We established a TD risk prediction method based on the parameters of the machine-learned ranking model. There are several objective function designs for ranking models [32,33]. In particular, pointwise [26], pairwise [34-36], and listwise [37,38] approaches have been proposed. Furthermore, several learning algorithms have been developed, including ones that use logistic regression, neural networks [39], and boosting [40].

We designed the model on the basis of the pairwise approach and used logistic regression. The pairwise approach was appropriate as the only rating scale for learning was the TD risk score. Logistic regression was selected because it was the most frequently used approach in related work [24] and because it was used in our previous work [25].

We hypothesized that the risk of TD of patient p_m can be calculated from a feature vector x_m that incorporates a variety of patient information up to time t_m . Therefore, we assumed that the scalar TD risk can be represented by the inner product of a weight vector and the feature vector, that is, $w \cdot x_m$. To obtain the weight vector w , we modeled the probability that patient p_m at time t_m would discontinue treatment earlier than p_n at t_n , with x_m and x_n attributed to $y_{m,n}$ with the logistic regression:

$$P(y_{m,n} | x_m, x_n; w) = 1 / \{1 + \exp[-y_{m,n} w(x_m - x_n)]\}$$

The notation $w(x_m - x_n)$ denotes the scalar product of w and $x_m - x_n$.

ML Design

The ranking method, based on the pairwise approach, requires pairs of data for optimizing the parameters of the model. In general, $n(n-1)/2$ pairs can be generated for n records with no censoring. As this study included censored data that were TCs, all pairs for optimization must satisfy the abovementioned combination rule. There was also a concern that the model would have a heavier bias toward TD cases than toward TC cases. According to survey papers [41-43] on biased data, sampling has often been attempted as a way to solve this problem [44,45]. We took the means of sampling 1 record from each patient to prevent biased learning on a small number of patients. When the w estimate was computed, we randomly selected 1 recorded

date of a hospital visit for each patient and used the date t_m or t_n as the starting point of TD or TC to calculate TD (p_m, t_m) or TC (p_n, t_n). The number of all pairs satisfying the abovementioned combination rule with the sampling was 867,574 in the training data and 17,038 in the test data. The computational complexity of pairwise-based ranking learning is $O(n^2)$. The sampling results in a slightly reduced computational cost.

When the training data size, N , is smaller than the dimension of the feature vectors, or when sampling of the training data is biased, a maximum-likelihood estimation often overfits a logistic regression model to the training data, leading the model to rank many new patients inaccurately. We used an L2-norm regularization method [23] to mitigate overfitting and improve the generalizability of the model, as we did in our previous study [25].

Using training data $[(x_1, x_2, y_{1,2}), \dots, (x_1, x_N, y_{1,N}), \dots, (x_2, x_3, y_{2,3}), \dots, (x_m, x_n, y_{m,n}), \dots, (x_{N-1}, x_N, y_{N-1,N})]$, we estimated w as follows:

$$\begin{aligned} \hat{w} &= \arg \max \left\{ \sum_{n=1}^N \log P(y_{m,n} | x_m, x_n; w) - \lambda \|w\|_2^2 \right\} \\ &= \arg \max \left\{ \sum_{n=1}^N \log(1 / (1 + \exp(-y_{m,n} w(x_m - x_n)))) - \lambda \|w\|_2^2 \right\} \end{aligned}$$

where the squared L2-norm of w , $\|w\|_2^2$, is an L2-norm regularizer that acts as a mitigating penalty to provide large absolute weight values only to frequently occurring features in the training data.

The symbol λ is a hyperparameter for regularization and was tuned as follows: the training data were randomly split into 2 sets of data and used in a 2-fold cross-validation test; for each test, the prediction accuracy was evaluated with one set of data for training and the other set of data for testing, with λ set to 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, and 100. The value of λ at which

the average prediction accuracy of the 2 tests was highest was chosen.

TD Risk Score Design

The TD risk score of patient p_m at time t_m is represented by the logit value $w \cdot x_m$. The higher the value of the TD risk, the earlier TD is predicted to occur. Figure 2 shows an example of the TD risk value.

Statistical Analysis

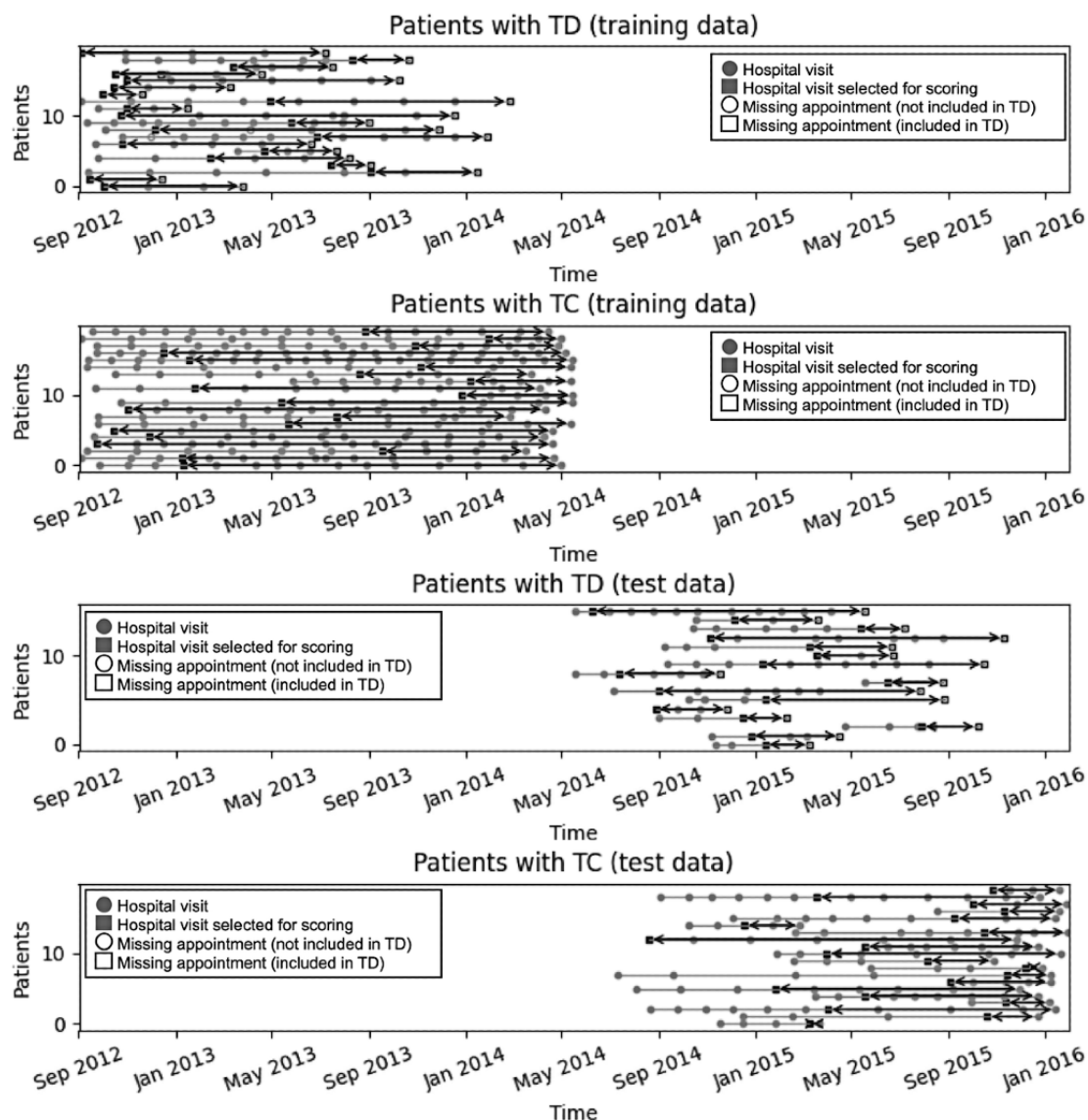
We implemented the model and ML optimization in-house in C and Python 3.7 and used it in all the experiments.

Results

Distribution of TD and TC

The detailed demographic data are shown in Table 1. The average numbers of appointments by patients with TD and TC were 4.8 and 10.4, respectively, in the training group and 3.1 and 5.8, respectively, in the testing group. The difference in distribution was because of the training and test data were classified according to whether or not they had a history of hospital visits before May 17, 2014, and the duration of the training data (828 days) was approximately twice that of the test data (415 days). Furthermore, as shown in Figure 3, the training data included patients who had been attending the hospital since before September 3, 2012, which was the starting point for the experiment; thus, patients with TC in the training data tended to have more appointments. In contrast, patients with TC in the test data tended to have fewer appointments, as these data were limited to patients who had attended the hospital since May 17, 2014. However, the number of appointments for patients with TD was low for both training and test data as patients with TD generally had shorter hospital visits. The average numbers of MAs by patients with TD and TC were 1.6 and 1.6, respectively, in the training group and 1.2 and 1.3, respectively, in the testing group.

Figure 3. Example of distribution of visit and appointment dates. TC: treatment continuation; TD: treatment discontinuation.



Predictive Performance Against TD

The hyperparameter λ of the machine-learned ranking model was tuned with 2 cross-validations, and it was set to 10 in the testing stage. The C-index of the predicted ranking was calculated as the number of correctly ranked pairs divided by the total number of comparable pairs. During testing, the TD risk score generated by the algorithm performed well, with a C-index (95% confidence limits) of 0.749 (0.655, 0.823), and outperformed the Cox regression model, with a C-index (95% confidence limits) of 0.662 (0.574, 0.748). As shown by the Kaplan-Meier curve in Figure 4, it was able to correctly model the population at high risk for TD. 10.3% (36/349) of the patients whose calibrated risk scores were ≥ 0.5 discontinued treatment within 100 days, whereas 93.9% (651/693) of the patients whose scores were < 0.5 continued treatment for over 1 year.

The number of TD cases was much smaller in the data used in this study than the number of patients who did not interrupt their visits. As validation with the C-index alone might not be

sufficient to evaluate the performance in the case of imbalanced data [45,46], the AUPRC was used in addition to the AUROC to evaluate whether the risk score could predict TD in a specific period, as shown in Table 3. Both the AUROC and AUPRC of the TD risk score were higher than those of the Cox regression model.

TD prediction within 6 months showed an AUROC (95% confidence limits) of 0.741 (0.641, 0.833) and an AUPRC (95% confidence limits) of 0.335 (0.193, 0.499). These values at 1 year were 0.758 (0.649, 0.857) and 0.713 (0.554, 0.841), respectively.

Subsequently, the TD risk score was converted to a range of 0 to 1 to validate the performance of risk stratification. As shown in the calibration plot using the test data in Figure 5, the observed and predicted TD rates were relatively correlated. These results indicate that the TD risk score can provide clinicians with information about the risk of TD in advance with favorable predictive performance and improve patient outcomes by providing room for interventions to avoid interruptions.

Figure 4. Kaplan-Meier curves displaying the probability of treatment discontinuation (TD) for the 2 groups of test data divided by the median TD risk scores obtained from the training data.

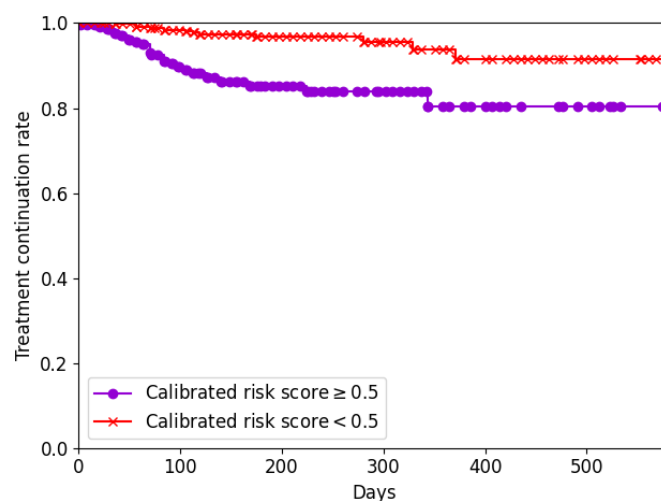


Table 3. Predictive performance against TD^a.

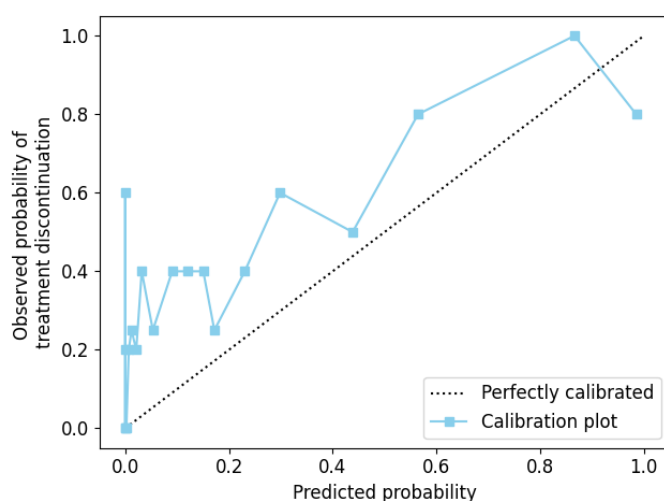
Months	AUROC ^b , mean (95% confidence limits)		AUPRC ^c , mean (95% confidence limits)	
	Ranking model	Cox model	Ranking model	Cox model
2	0.747 (0.607, 0.868)	0.668 (0.544, 0.787)	0.081 (0.024, 0.299)	0.035 (0.016, 0.071)
3	0.776 (0.666, 0.870)	0.691 (0.581, 0.793)	0.228 (0.090, 0.412)	0.136 (0.052, 0.262)
4	0.748 (0.637, 0.844)	0.641 (0.531, 0.746)	0.290 (0.139, 0.470)	0.156 (0.072, 0.278)
5	0.751 (0.651, 0.843)	0.666 (0.557, 0.768)	0.309 (0.163, 0.483)	0.215 (0.107, 0.360)
6	0.741 (0.641, 0.833)	0.645 (0.533, 0.751)	0.335 (0.193, 0.499)	0.236 (0.127, 0.379)
7	0.746 (0.645, 0.841)	0.660 (0.547, 0.764)	0.414 (0.254, 0.576)	0.308 (0.172, 0.468)
8	0.752 (0.650, 0.846)	0.677 (0.565, 0.781)	0.478 (0.311, 0.635)	0.384 (0.227, 0.544)
9	0.756 (0.654, 0.850)	0.675 (0.561, 0.785)	0.510 (0.337, 0.670)	0.438 (0.269, 0.601)
10	0.750 (0.646, 0.846)	0.691 (0.569, 0.800)	0.570 (0.402, 0.726)	0.562 (0.389, 0.708)
11	0.732 (0.625, 0.830)	0.680 (0.561, 0.793)	0.609 (0.442, 0.757)	0.597 (0.426, 0.742)
12	0.758 (0.649, 0.857)	0.687 (0.569, 0.798)	0.713 (0.554, 0.841)	0.645 (0.485, 0.784)

^aTD: treatment discontinuation.

^bAUROC: area under the receiver operating characteristic curve.

^cAUPRC: area under the precision-recall curve.

Figure 5. The distribution of the predicted probability and observed probability of treatment discontinuation is shown in a line chart. Each point represents the observed and predicted probabilities for each of the 20 segments of the test population.



Items With the Largest Coefficient Values

The items with the largest coefficient values were examined to check for leakage, wherein unintended information is used for prediction and degrades the performance of the model. The 5

highest and the 5 lowest items are shown in [Table 4](#). The specific mechanism by which each item contributes to the prediction is difficult to discuss at this time, but there were no items among the top 5 that suggested obvious leakage.

Table 4. Top 5 and bottom 5 explanatory variables obtained from the training set.

Category	Weight size	Feature
Top 1	8.1	Frequency of visits with the reservation at the department of cardiovascular medicine within 3 months
Top 2	5.2	Frequency of visits with no letter of reference within 6 months
Top 3	5.2	Frequency of visits with no letter of reference within 3 months
Top 4	5.2	Frequency of visits with the reservation before an operation in the department of cardiovascular medicine
Top 5	5.2	Frequency of laboratory tests of protein in urine within 6 months
Bottom 1	-28	Frequency of blood pressure tests within 3 months
Bottom 2	-25	Frequency of appointments of carotid artery ultrasound examination within 3 months
Bottom 3	-16	Frequency of carotid echo tests within 3 months
Bottom 4	-15	Frequency of laboratory tests of HbA _{1c} ^a within 6 months
Bottom 5	-15	Frequency of laboratory tests of HbA _{1c} within 1 year

^aHbA_{1c}: hemoglobin A_{1c}.

Discussion

Principal Findings

In this study, we generated a prediction model for the risk of TD using approximately 150,000 explanatory variables extracted from EHRs and advanced machine-learned techniques. The accuracy of the model's prediction was validated.

Comparison With Prior Work

ML has been used in almost all aspects of diabetic research, especially in biomarker identification and diagnosis prediction [47-50]. The prediction of interruptions in medical visits requires the use of survival time analysis to build a model. However, there are few studies that have used ML for this purpose. In our study, to avoid the proportional hazard assumption of the Cox

regression model and learning difficulties because of imbalanced data, we implemented a ranking method and showed that the scores calculated for each patient using the parameters obtained from the training data were useful for predicting TD, as shown in [Table 3](#).

Our method is a novel way of constructing a survival regression model, and our experimental evaluation showed that it outperformed the existing Cox model in terms of the C-index and AUROC and AUPRC measures and that it would be a useful option for imbalanced data such as TD. The obtained level of performance was not significantly superior to that of the Cox regression model with regard to CIs. Nonetheless, it was not inferior. Many prediction tasks in the clinical domain require that imbalanced data be addressed by prediction models using survival time analysis. Our modeling method does not require

the proportional hazard assumption of the Cox regression model and avoids the problem of learning from imbalanced data. It has no variable assumptions, which allowed us to use approximately 150,000 features. Therefore, we believe that our method is a new option for survival regression models in the clinical field.

Limitations

Our study had several important limitations that must be mentioned. First, the data were obtained from just one hospital. In addition, the test data were obtained by splitting up the data from just one hospital. They may not be entirely representative of other regions because of the different implementations and degrees of diabetes care. Consequently, the results of this study are not sufficient to assess the generalizability of our method; a study using more data from different hospitals will be required.

Second, the participants with a history of TD in this study represented only 1 subgroup of patients. Some could have discontinued treatment temporarily, and we were unable to capture these patients in this study. Moreover, if a patient changed clinics without notice and continued treatment elsewhere without any evidence in the EHR, their case would have been judged as TD cases, even if that would not have been accurate. Nonetheless, because this study relied on EHR information, the findings serve the purpose of evaluating the accuracy of the model using real-world data.

Third, our method used a large number of features and optimized them with the L2-norm regularizer, which made it difficult to find features of high importance that contribute to the prediction. In the future, we intend to investigate ways to improve

interpretability, such as by using explainable artificial intelligence and Lasso regularization.

Fourth, a large number of features were generated in the predefined procedure, and the inherent trends and meanings of each feature in itself are not adequately considered. The features need to be designed more appropriately to improve the interpretability of the results.

Fifth, our method was superior to the binary classification model in that it could compare a patient's risk of TD with the time remaining until TD. However, it requires $O(n^2)$ pairs to learn the model parameters, whereas a binary classification requires only $O(n)$ records for n training data. We need to reduce the computational cost.

Finally, it should be noted that as ML generally reflects the characteristics of the majority, our results suggest that the predictive performance obtained in this study cannot be applied to a minority of clusters in the population, such as pediatric patients.

Conclusions

We developed a novel prediction model for calculating the TD risk score by applying a machine-learned ranking model to EHR data. This score showed high prediction performance and outperformed the Cox regression model. Our model can alert clinicians about the risk of TD in advance and would be useful in improving patient outcomes by providing room for interventions to avoid interruptions and support diabetes care. In addition to estimating the TD risk score, we are studying ways to predict glycemic control in patients with diabetes to further improve their care.

Acknowledgments

This work was funded by the University of Tokyo and Nippon Telegraph and Telephone Corporation in a joint research program that was conducted at the University of Tokyo Center of Innovation, Sustainable Life Care, and the Ageless Society and dedicated to self-managing health care in the Aging Society of Japan. The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Tokyo Center of Innovation.

Data Availability

The data in this study are not openly available because of the restrictions imposed by the research ethics committees that approved this study.

Conflicts of Interest

HK, KH, and AF are employees of the Nippon Telegraph and Telephone Corporation (NTT), Tokyo, Japan. AC was an employee of NTT and is now an employee of NTT DOCOMO, Inc, Tokyo, Japan. TH was an employee of NTT and is now the chief executive officer of the NTT-AT IPS Corporation, Kanagawa, Japan.

References

1. Diabetes Control and Complications Trial Research Group. Effect of intensive diabetes treatment on the development and progression of long-term complications in adolescents with insulin-dependent diabetes mellitus: diabetes control and complications trial. *J Pediatrics* 1994 Aug;125(2):177-188. [doi: [10.1016/s0022-3476\(94\)70190-3](https://doi.org/10.1016/s0022-3476(94)70190-3)] [Medline: [8040759](https://pubmed.ncbi.nlm.nih.gov/8040759/)]
2. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ* 2000 Aug 12;321(7258):405-412 [FREE Full text] [doi: [10.1136/bmj.321.7258.405](https://doi.org/10.1136/bmj.321.7258.405)] [Medline: [10938048](https://pubmed.ncbi.nlm.nih.gov/10938048/)]

3. Archibald LK, Gill GV. Diabetic clinic defaulters — who are they and why do they default? *Pract Diab Int* 1992 Jan;9(1):13-14. [doi: [10.1002/pdi.1960090104](https://doi.org/10.1002/pdi.1960090104)]
4. Hammersley MS, Holland MR, Walford S, Thorn PA. What happens to defaulters from a diabetic clinic? *Br Med J (Clin Res Ed)* 1985 Nov 09;291(6505):1330-1332 [FREE Full text] [doi: [10.1136/bmj.291.6505.1330](https://doi.org/10.1136/bmj.291.6505.1330)] [Medline: [3933654](https://pubmed.ncbi.nlm.nih.gov/3933654/)]
5. American Diabetes Association. Standards of medical care in diabetes—2020 abridged for primary care providers. *Clin Diabetes* 2020 Jan;38(1):10-38 [FREE Full text] [doi: [10.2337/cd20-as01](https://doi.org/10.2337/cd20-as01)] [Medline: [31975748](https://pubmed.ncbi.nlm.nih.gov/31975748/)]
6. Currie C, Peyrot M, Morgan C, Poole CD, Jenkins-Jones S, Rubin RR, et al. The impact of treatment noncompliance on mortality in people with type 2 diabetes. *Diabetes Care* 2012 Jun;35(6):1279-1284 [FREE Full text] [doi: [10.2337/dc11-1277](https://doi.org/10.2337/dc11-1277)] [Medline: [22511257](https://pubmed.ncbi.nlm.nih.gov/22511257/)]
7. Graber A, Davidson P, Brown A, McRae J, Woolridge K. Dropout and relapse during diabetes care. *Diabetes Care* 1992 Nov;15(11):1477-1483. [doi: [10.2337/diacare.15.11.1477](https://doi.org/10.2337/diacare.15.11.1477)] [Medline: [1468274](https://pubmed.ncbi.nlm.nih.gov/1468274/)]
8. Gucciardi E, Demelo M, Offenheim A, Stewart DE. Factors contributing to attrition behavior in diabetes self-management programs: a mixed method approach. *BMC Health Serv Res* 2008 Feb 04;8:33 [FREE Full text] [doi: [10.1186/1472-6963-8-33](https://doi.org/10.1186/1472-6963-8-33)] [Medline: [18248673](https://pubmed.ncbi.nlm.nih.gov/18248673/)]
9. Kawahara R, Amemiya T, Yoshino M, Miyamae M, Sasamoto K, Omori Y. Dropout of young non-insulin-dependent diabetics from diabetic care. *Diabetes Res Clin Pract* 1994 Jul;24(3):181-185. [doi: [10.1016/0168-8227\(94\)90114-7](https://doi.org/10.1016/0168-8227(94)90114-7)]
10. Sone H, Kawai K, Takagi H, Yamada N, Kobayashi M. Outcome of one-year of specialist care of patients with type 2 diabetes: a multi-center prospective survey (JDDM 2). *Intern Med* 2006;45(9):589-597 [FREE Full text] [doi: [10.2169/internalmedicine.45.1609](https://doi.org/10.2169/internalmedicine.45.1609)] [Medline: [16755089](https://pubmed.ncbi.nlm.nih.gov/16755089/)]
11. Noda M, Yamazaki K, Hayashino Y, Izumi K, Goto A. Japanese practice guidance to improve patients' adherence to appointments for diabetes care. *Human Data*. 2019 Jul 15. URL: https://human-data.or.jp/wp/wp-content/uploads/2018/07/dm_jushinchudan_guide43_e.pdf [accessed 2022-01-31]
12. Lee RR, Samsudin MI, Thirumoorthy T, Low LL, Kwan YH. Factors affecting follow-up non-attendance in patients with Type 2 diabetes mellitus and hypertension: a systematic review. *Singapore Med J* 2019 May;60(5):216-223 [FREE Full text] [doi: [10.11622/smedj.2019042](https://doi.org/10.11622/smedj.2019042)] [Medline: [31187148](https://pubmed.ncbi.nlm.nih.gov/31187148/)]
13. Masuda Y, Kubo A, Kokaze A, Yoshida M, Sekiguchi K, Fukuhara N, et al. Personal features and dropout from diabetic care. *Environ Health Prev Med* 2006 May;11(3):115-119 [FREE Full text] [doi: [10.1265/ehpm.11.115](https://doi.org/10.1265/ehpm.11.115)] [Medline: [21432385](https://pubmed.ncbi.nlm.nih.gov/21432385/)]
14. Benoit SR, Ji M, Fleming R, Philis-Tsimikas A. Predictors of dropouts from a San Diego diabetes program: a case control study. *Prev Chronic Dis* 2004 Oct;1(4):A10 [FREE Full text] [Medline: [15670442](https://pubmed.ncbi.nlm.nih.gov/15670442/)]
15. Karter AJ, Parker MM, Moffet HH, Ahmed AT, Ferrara A, Liu JY, et al. Missed appointments and poor glycemic control: an opportunity to identify high-risk diabetic patients. *Med Care* 2004 Feb;42(2):110-115. [doi: [10.1097/01.mlr.0000109023.64650.73](https://doi.org/10.1097/01.mlr.0000109023.64650.73)] [Medline: [14734947](https://pubmed.ncbi.nlm.nih.gov/14734947/)]
16. Díaz EG, Medina DR, López AG, Porras M. Determinants of adherence to hypoglycemic agents and medical visits in patients with type 2 diabetes mellitus. *Endocrinol Diabetes Nutr* 2017 Dec;64(10):531-538. [doi: [10.1016/j.endinu.2017.08.004](https://doi.org/10.1016/j.endinu.2017.08.004)] [Medline: [29108925](https://pubmed.ncbi.nlm.nih.gov/29108925/)]
17. Rhee MK, Slocum W, Ziemer DC, Culler SD, Cook CB, El-Kebbi IM, et al. Patient adherence improves glycemic control. *Diabetes Educ* 2005;31(2):240-250. [doi: [10.1177/0145721705274927](https://doi.org/10.1177/0145721705274927)] [Medline: [15797853](https://pubmed.ncbi.nlm.nih.gov/15797853/)]
18. Fullerton B, Erler A, Pöhlmann B, Gerlach FM. Predictors of dropout in the German disease management program for type 2 diabetes. *BMC Health Serv Res* 2012 Jan 10;12(1):8 [FREE Full text] [doi: [10.1186/1472-6963-12-8](https://doi.org/10.1186/1472-6963-12-8)] [Medline: [22233930](https://pubmed.ncbi.nlm.nih.gov/22233930/)]
19. Buys KC, Selleck C, Buys DR. Assessing retention in a free diabetes clinic. *J Nurse Practitioners* 2019 Apr;15(4):301-5.e1. [doi: [10.1016/j.nurpra.2018.12.003](https://doi.org/10.1016/j.nurpra.2018.12.003)]
20. Wong M, Haswell-Elkins M, Tamwoy E, McDermott R, d'Abbs P. Perspectives on clinic attendance, medication and foot-care among people with diabetes in the Torres Strait Islands and Northern Peninsula Area. *Aust J Rural Health* 2005 Jun;13(3):172-177. [doi: [10.1111/j.1440-1854.2005.00678.x](https://doi.org/10.1111/j.1440-1854.2005.00678.x)] [Medline: [15932487](https://pubmed.ncbi.nlm.nih.gov/15932487/)]
21. Gibson DM. Frequency and predictors of missed visits to primary care and eye care providers for annually recommended diabetes preventive care services over a two-year period among U.S. adults with diabetes. *Prev Med* 2017 Dec;105:257-264. [doi: [10.1016/j.ypmed.2017.09.019](https://doi.org/10.1016/j.ypmed.2017.09.019)] [Medline: [28963006](https://pubmed.ncbi.nlm.nih.gov/28963006/)]
22. Sun C, Taylor K, Levin S, Renda SM, Han H. Factors associated with missed appointments by adults with type 2 diabetes mellitus: a systematic review. *BMJ Open Diabetes Res Care* 2021 Mar 05;9(1):e001819 [FREE Full text] [doi: [10.1136/bmjdr-2020-001819](https://doi.org/10.1136/bmjdr-2020-001819)] [Medline: [33674280](https://pubmed.ncbi.nlm.nih.gov/33674280/)]
23. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY: Springer; 2006.
24. Carreras-García D, Delgado-Gómez D, Llorente-Fernández F, Arribas-Gil A. Patient no-show prediction: a systematic literature review. *Entropy (Basel)* 2020 Jun 17;22(6):675 [FREE Full text] [doi: [10.3390/e22060675](https://doi.org/10.3390/e22060675)] [Medline: [33286447](https://pubmed.ncbi.nlm.nih.gov/33286447/)]
25. Kurasawa H, Hayashi K, Fujino A, Takasugi K, Haga T, Waki K, et al. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. *J Diabetes Sci Technol* 2016 May;10(3):730-736 [FREE Full text] [doi: [10.1177/1932296815614866](https://doi.org/10.1177/1932296815614866)] [Medline: [26555782](https://pubmed.ncbi.nlm.nih.gov/26555782/)]
26. Liu T. Learning to rank for information retrieval. *FNT Inform Retrieval* 2009;3(3):225-331. [doi: [10.1561/1500000016](https://doi.org/10.1561/1500000016)]
27. Wang P, Li Y, Reddy CK. Machine learning for survival analysis. *ACM Comput Surv* 2019 Nov 30;51(6):1-36. [doi: [10.1145/3214306](https://doi.org/10.1145/3214306)]

28. Cox DR. Regression models and life-tables. *J Royal Statistical Soc Series B (Methodological)* 2018 Dec 05;34(2):187-202. [doi: [10.1111/j.2517-6161.1972.tb00899.x](https://doi.org/10.1111/j.2517-6161.1972.tb00899.x)]
29. Raykar V, Steck H, Krishnapuram B, Dehing-Oberije C, Lambin P. On ranking in survival analysis: bounds on the concordance index. In: *Proceedings of the Advances in Neural Information Processing Systems 20 (NIPS 2007)*. 2007 Presented at: *Advances in Neural Information Processing Systems 20 (NIPS 2007)*; Dec 3-6, 2007; Vancouver, British Columbia. [doi: [10.5555/2981562.2981714](https://doi.org/10.5555/2981562.2981714)]
30. Van Belle V, Pelckmans K, Van Huffel S, Suykens JA. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif Intell Med* 2011 Oct;53(2):107-118. [doi: [10.1016/j.artmed.2011.06.006](https://doi.org/10.1016/j.artmed.2011.06.006)] [Medline: [21821401](https://pubmed.ncbi.nlm.nih.gov/21821401/)]
31. Chen H, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. *BMC Med Res Methodol* 2012 Jul 23;12:102 [FREE Full text] [doi: [10.1186/1471-2288-12-102](https://doi.org/10.1186/1471-2288-12-102)] [Medline: [22824262](https://pubmed.ncbi.nlm.nih.gov/22824262/)]
32. Burges C, Ragno R, Le Q. Learning to rank with nonsmooth cost functions. In: *Advances in Neural Information Processing Systems 19*. Cambridge, Massachusetts, United States: MIT Press; 2006.
33. Donmez P, Svore K, Burges CJ. On the local optimality of LambdaRank. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009 Presented at: *SIGIR '09: The 32nd International ACM SIGIR conference on research and development in Information Retrieval*; Jul 19 - 23, 2009; Boston MA USA. [doi: [10.1145/1571941.1572021](https://doi.org/10.1145/1571941.1572021)]
34. Cao Z, Qin T, Liu T, Tsai M, Li H. Learning to rank: from pairwise approach to listwise approach. In: *Proceedings of the 24th international conference on Machine learning*. 2007 Presented at: *ICML '07 & ILP '07: The 24th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming*; Jun 20 - 24, 2007; Corvallis Oregon USA. [doi: [10.1145/1273496.1273513](https://doi.org/10.1145/1273496.1273513)]
35. Furnkranz J, Hullermeier E. Preference learning and ranking by pairwise comparison. In: *Preference Learning*. Berlin, Heidelberg: Springer; 2010.
36. Usunier N, Buffoni D, Gallinari P. Ranking with ordered weighted pairwise classification. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009 Presented at: *ICML '09: The 26th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming*; Jun 14 - 18, 2009; Montreal Quebec Canada. [doi: [10.1145/1553374.1553509](https://doi.org/10.1145/1553374.1553509)]
37. Xia F, Liu T, Wang J, Zhang W, Li H. Listwise approach to learning to rank: theory and algorithm. In: *Proceedings of the 25th international conference on Machine learning*. 2008 Presented at: *ICML '08: The 25th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming*; Jul 5 - 9, 2008; Helsinki Finland. [doi: [10.1145/1390156.1390306](https://doi.org/10.1145/1390156.1390306)]
38. Shi Y, Larson M, Hanjalic A. List-wise learning to rank with matrix factorization for collaborative filtering. In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010 Presented at: *RecSys '10: Fourth ACM Conference on Recommender Systems*; Sep 26 - 30, 2010; Barcelona Spain. [doi: [10.1145/1864708.1864764](https://doi.org/10.1145/1864708.1864764)]
39. Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, et al. Learning to rank using gradient descent. In: *Proceedings of the 22nd international conference on Machine learning*. 2005 Presented at: *ICML '05: Proceedings of the 22nd international conference on Machine learning*; Aug 7 - 11, 2005; Bonn Germany. [doi: [10.1145/1102351.1102363](https://doi.org/10.1145/1102351.1102363)]
40. Freund Y, Iyer R, Schapire R, Singer Y. An efficient boosting algorithm for combining preferences. *J Mach Learn Res* 2003;4:933-969. [doi: [10.5555/945365.964285](https://doi.org/10.5555/945365.964285)]
41. He H, Garcia E. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009 Sep;21(9):1263-1284. [doi: [10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239)]
42. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. *Int J Patt Recogn Artif Intell* 2011 Nov 21;23(04):687-719. [doi: [10.1142/S0218001409007326](https://doi.org/10.1142/S0218001409007326)]
43. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. *Intelligent Data Analysis* 2002 Nov 15;6(5):429-449. [doi: [10.3233/ida-2002-6504](https://doi.org/10.3233/ida-2002-6504)]
44. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. *ACM Comput Surv* 2016 Nov 11;49(2):1-50. [doi: [10.1145/2907070](https://doi.org/10.1145/2907070)]
45. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inform Sci* 2013 Nov;250:113-141. [doi: [10.1016/j.ins.2013.07.007](https://doi.org/10.1016/j.ins.2013.07.007)]
46. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10(3):e0118432 [FREE Full text] [doi: [10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432)] [Medline: [25738806](https://pubmed.ncbi.nlm.nih.gov/25738806/)]
47. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016 Dec 13;316(22):2402-2410. [doi: [10.1001/jama.2016.17216](https://doi.org/10.1001/jama.2016.17216)] [Medline: [27898976](https://pubmed.ncbi.nlm.nih.gov/27898976/)]
48. 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](https://doi.org/10.1016/j.csbj.2016.12.005)] [Medline: [28138367](https://pubmed.ncbi.nlm.nih.gov/28138367/)]

49. Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *J Diabetes Sci Technol* 2015 Jan;9(1):86-90 [FREE Full text] [doi: [10.1177/1932296814554260](https://doi.org/10.1177/1932296814554260)] [Medline: [25316712](https://pubmed.ncbi.nlm.nih.gov/25316712/)]
50. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int J Med Inform* 2017 Jan;97:120-127 [FREE Full text] [doi: [10.1016/j.ijmedinf.2016.09.014](https://doi.org/10.1016/j.ijmedinf.2016.09.014)] [Medline: [27919371](https://pubmed.ncbi.nlm.nih.gov/27919371/)]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
EHR: electronic health record
HbA_{1c}: hemoglobin A_{1c}
MA: missed appointment
ML: machine learning
NTT: Nippon Telegraph and Telephone Corporation
TC: treatment continuation
TD: treatment discontinuation

Edited by A Mavragani; submitted 28.03.22; peer-reviewed by R Bellazzi, G Nneji, G Lim; comments to author 29.05.22; revised version received 19.06.22; accepted 02.09.22; published 23.09.22

Please cite as:

Kurasawa H, Waki K, Chiba A, Seki T, Hayashi K, Fujino A, Haga T, Noguchi T, Ohe K
Treatment Discontinuation Prediction in Patients With Diabetes Using a Ranking Model: Machine Learning Model Development
JMIR Bioinform Biotech 2022;3(1):e37951
URL: <https://bioinform.jmir.org/2022/1/e37951>
doi: [10.2196/37951](https://doi.org/10.2196/37951)
PMID:

©Hisashi Kurasawa, Kayo Waki, Akihiro Chiba, Tomohisa Seki, Katsuyoshi Hayashi, Akinori Fujino, Tsuneyuki Haga, Takashi Noguchi, Kazuhiko Ohe. Originally published in *JMIR Bioinformatics and Biotechnology* (<https://bioinform.jmir.org>), 23.09.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Bioinformatics and Biotechnology*, is properly cited. The complete bibliographic information, a link to the original publication on <https://bioinform.jmir.org/>, as well as this copyright and license information must be included.