Prediction model of Graves’ disease in general clinical practice based on complete blood count and biochemistry profile

Ai Yoshihara1), Jaeduk Yoshimura Noh1), Kosuke Inoue1),2), Junichi Taguchi3), Keisuke Hata4), Toru Aizawa4), Yoshino Taira Arai4), Natsuko Watanabe4), Miho Fukushita4), Masako Matsumoto1), Nami Suzuki1), Ayako Hoshiyama1), Ai Suzuki1), Takako Mitsumatsu1), Aya Kinoshita1), Kentaro Mikura1), Ran Yoshimura1), Kiminori Sugino1) and Koichi Ito1)

1) Ito Hospital, Tokyo 150-8308, Japan
2) Department of Social Epidemiology, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan
3) Tokyo Midtown Clinic, Tokyo 107-6206, Japan
4) Nihonbashi Muromachi Mitsui Tower Midtown Clinic, Tokyo 103-0022, Japan

Abstract. Although untreated Graves’ disease (GD) is associated with a higher risk of cardiac complications and mortality, there is no well-established way to predict the onset of thyrotoxicosis in clinical practice. The aim of this study was to identify important variables that will make it possible to predict GD and thyrotoxicosis (GD + painless thyroiditis (PT)) by using a machine-learning-based model based on complete blood count and standard biochemistry profile data. We identified 19,335 newly diagnosed GD patients, 3,267 PT patients, and 4,159 subjects without any thyroid disease. We built a GD prediction model based on information obtained from subjects regarding sex, age, a complete blood count, and a standard biochemistry profile. We built the model in the training set and evaluated the performance of the model in the test set by using the artificial intelligence software Prediction One. Our machine learning-based model showed high discriminative ability to predict GD in the test set (area under the curve [AUC] 0.99). The main contributing factors to predict GD included age and serum creatinine, total cholesterol, alkaline phosphatase, and total protein levels. We still found high discriminative ability even when we restricted the variables to these five most contributory factors in our prediction model (AUC 0.97) built by using artificial intelligence software showed high GD prediction ability based on information regarding only five factors.

Key words: Prediction model, Thyrotoxicosis, Standard biochemistry profile, Machine learning

SINCE THE SYMPTOMS of thyrotoxicosis vary from person to person, the diagnosis of thyrotoxicosis may be delayed because thyroid function tests are not routinely performed during medical checkups. Graves’ disease (GD) and painless thyroiditis (PT) are the main causes of thyrotoxicosis [1]. The diagnosis of GD is confirmed by laboratory findings, including by elevated serum FT4 and FT3 levels, suppression of TSH, and the presence of thyroid receptor antibody (TRAb) and/or increased radioactive iodine uptake (RAIU) [2]. Early diagnosis of GD may be of benefit in preventing severe cardiac or musculoskeletal complications [3-6]. PT is caused by destruction of thyroid tissue and is characterized by transient thyrotoxicosis followed by transient hypothyroidism and decreased radioactive iodine uptake, and the thyrotoxicosis that develops in PT patients mostly resolves without medication. Medical checkups are usually performed annually in Japan, and being able to predict thyrotoxicosis based on blood test and biochemistry parameters would be helpful in allowing general practitioners to diagnose thyroid dysfunction in the early stage. Symptoms of thyrotoxicosis are weight loss, muscle weakness, palpitations, tremor, heat intolerance, and persistent fatigue. Ophthalmopathy and goiter are the signs of GD. Diagnosing thyrotoxicosis may be simple for endocrinologists, but the symptoms are not specific to thyroid disorders, and general practitioners may not suspect thyrotoxicosis and fail to measure thyroid hormone levels even when patients complain of such symptoms [7, 8]. Thus, it is important to identify key variables that will enable prediction of GD based on commonly
reported or measured information so that general physicians do not miss GD and order thyroid hormone level measurements in a timely manner. In addition, thyroid function tests are expensive, and if thyroid disorders can be predicted by a general blood test, it will help reduce medical costs.

In this study we therefore built a prediction model for GD and thyrotoxicosis (PT and GD) based on a complete blood count and standard biochemistry profile by using the artificial intelligence software Prediction One (Sony Network Communications Inc., Tokyo, Japan) and conventional logistic regression.

Methods

Study samples

We identified all newly diagnosed GD patients and PT patients who made their first visit to our hospital between January 1, 2005 and December 31, 2018. We excluded patients who had been treated for thyroid disorders before their first visit to our hospital and patients who had been taking medication that might affect thyroid function. We also collected euthyroid subjects with no thyroid disorders who made their first visit to our hospital during the same period (control subjects). Control subjects (without any thyroid diseases) had a thyroid that showed a homogeneous echo pattern on thyroid images without any thyroid nodules, did not have a goiter, and tested negative for antithyroid antibodies (TRAb, thyroglobulin antibody, thyroid peroxidase antibody). All study protocols were approved by the Ethics Committee of Tokyo Midtown Clinic and informed consent was obtained in the form of opt-out.

Predictors

The predictors in our models were chosen from complete blood count and standard biochemistry profile data obtained from the subjects during their first visit to our hospital. They included patient age, sex, and the values of the following parameters: RBC, Hb, Ht, MCV, MCH, MCHC, Plt, WBC, Neu, Lym, Mo, Es, Ba, TP, T-Bil, AST, ALT, LDH, γ-GTP, ALP, ChE, CPK, CRE, UA, Na, K, Cl, Ca, P, and T-Chol.

Outcomes

The primary outcome was GD, and the diagnosis of GD was made on the basis of an elevated serum FT3 level and FT4 level, suppression of the serum TSH level, negative TRAb, and decreased RAIU. FT3 and FT4 levels were measured by electrochemiluminescence immunoassays (ECLIA) (Elecsys FT3 and ECLusys FT4, Roche Diagnostics GmbH, Basel, Switzerland; manufacturer’s reference limits: 2.2–4.3 pg/mL and 0.8–1.6 ng/dL respectively). The TSH level was measured by an ECLIA (Elecsys TSH; Roche Diagnostics GmbH, Basel, Switzerland; manufacturer’s reference limits: 0.2–4.5 mIU/L). Control subjects (without any thyroid diseases) were identified on the basis of normal thyroid function, i.e., FT3, FT4, and TSH levels within their reference ranges, negative test for any of the thyroid autoantibodies, homogeneous echo pattern on thyroid images, i.e., absence of thyroid nodules, and absence of goiter.

Statistical analysis

Differences between the GD group and the normal group, between the GD group and the PT group, and between the thyrotoxicosis (GD + PT) group and the normal group were analyzed by the Mann-Whitney test. The analysis was performed in February 2021. In the training set (70% random sample of the total study sample), we built prediction models by using Prediction One, the ensemble learning model of neural networks and gradient-boosted decision tree, based on the 23 predictors listed above (sex, age, complete blood count data, and standard biochemistry profile data). Prediction One automatically adjusted and optimized the variables and created the best prediction model by an artificial neural network with internal cross-validation. We also conducted a multivariable logistic regression analysis for the prediction of GD and created a model to predict GD in the test set.

In the test set (remaining 30% of the total study sample), we assessed prediction performance based on the area under the receiver operating characteristic (ROC) curve, positive predictive value, and accuracy by Prediction One. We then identified the strong predictors of GD among the normal subjects.

We also investigated whether the predictors differed according to sex and the severity of GD (classified on the basis of the FT4 level at the time of diagnosis: severe GD, FT4 ≥5 ng/dL; mild GD, FT4 <5 ng/dL) [9].

We also built the prediction model using the logistic regression instead of Prediction One, and compared the prediction performance between these two models (i.e., logistic regression-based model vs. machine learning-based model).

P-values <0.05 were considered significant. The statistical analysis was performed using the JMP v14.
software program (SAS Institute, Inc., Cary, NC) and R version 4.1.0.

Results

Between January 1, 2005, and December 31, 2018, we identified 19,335 newly diagnosed GD patients, 3,267 PT patients, and 4159 subjects without any thyroid disease. Mean (±standard deviation; SD) age in the GD group, PT group, and normal group was 39.9 (±14.1) years, 39.5 (±14.2) years, and 34.3 (±13.4) years, respectively. The characteristics of the GD patients, PT patients, thyrotoxicosis patients, and normal subjects are shown in Table 1, Table 2, and Table 3.

Table 1  The characteristics of the GD patients and normal subjects

| Number | Graves Disease | Normal subjects | p value |
|--------|----------------|-----------------|---------|
| 19,335 | 4,159          |                 |         |
| Men:Female | 3,503:15,832 | 830:3,329 | < 0.001 |
| Age (yrs) | 39.9 (14.1) | 34.3 (13.4) | < 0.001 |
| RBC (×10^3/μL) | 465.1 (46.0) | 446.8 (42.7) | < 0.001 |
| Hb (g/dL) | 13.2 (1.3) | 13.6 (1.3) | < 0.001 |
| Ht (%) | 39.5 (3.7) | 40.3 (3.8) | < 0.001 |
| MCV (fL) | 85.2 (5.1) | 90.2 (5.0) | < 0.001 |
| MCH (pg) | 28.5 (2.0) | 30.4 (2.1) | < 0.001 |
| MCHC (%) | 33.5 (0.9) | 33.7 (0.8) | < 0.001 |
| Plt (×10^3/μL) | 24.4 (5.8) | 24.8 (5.5) | < 0.001 |
| WBC (/μL) | 5,824.8 (1,680.3) | 6,124.2 (1,643.4) | < 0.001 |
| Neu (%) | 50.6 (10.4) | 57.0 (9.3) | < 0.001 |
| Lym (%) | 36.3 (9.1) | 32.0 (8.1) | < 0.001 |
| Mo (%) | 10.0 (3.2) | 7.4 (2.1) | < 0.001 |
| Eo (%) | 2.6 (2.4) | 2.8 (2.5) | < 0.001 |
| Ba (%) | 0.6 (0.6) | 1.0 (0.5) | < 0.001 |
| TP (g/dL) | 6.9 (0.5) | 7.3 (0.4) | < 0.001 |
| T-Bil (mg/dL) | 0.7 (0.3) | 0.8 (0.3) | < 0.001 |
| AST (U/L) | 29.1 (15.9) | 20.6 (9.8) | < 0.001 |
| ALT (U/L) | 39.5 (30.4) | 19.5 (15.9) | < 0.001 |
| LDH (U/L) | 165.9 (29.9) | 168.8 (34.0) | < 0.001 |
| γGTP (U/L) | 39.2 (36.2) | 25.0 (48.1) | < 0.001 |
| ALP (U/L) | 355.2 (175.6) | 210.9 (128.5) | < 0.001 |
| ChE (U/L) | 387.3 (81.9) | 303.3 (72.8) | < 0.001 |
| CPK (U/L) | 62.3 (101.1) | 99.4 (171.1) | < 0.001 |
| CRE (mg/dL) | 0.5 (0.2) | 0.7 (0.1) | < 0.001 |
| UA (mg/dL) | 5.1 (1.2) | 4.6 (1.2) | < 0.001 |
| Na (mmol/L) | 140.0 (2.0) | 139.4 (1.9) | < 0.001 |
| K (mmol/L) | 4.3 (0.3) | 4.2 (0.3) | < 0.001 |
| Cl (mmol/L) | 105.3 (2.3) | 104.0 (2.3) | < 0.001 |
| Ca (mg/dL) | 9.6 (0.4) | 9.5 (0.4) | < 0.001 |
| P (mg/dL) | 4.0 (0.7) | 3.6 (0.6) | < 0.001 |
| T-C (mg/dL) | 148.0 (31.3) | 190.5 (35.5) | < 0.001 |
Prediction of GD among the normal subjects

The training dataset (70% random sample) included the data of 13,544 GD patients and 2,911 normal subjects, and the test dataset (30% random sample) included the data of 5,791 GD patients and 1,248 normal subjects. In the test set, our machine learning-based prediction model showed high discriminative ability as demonstrated by the ROC curves; the value of the area under the curve (AUC) was 0.99, positive predictive value 98.7%, accuracy 95.2%. The prediction model based on logistic regression also showed high discriminative ability for GD (AUC 0.98). The important predictors for the diagnosis of GD consisted of age and serum CRE, T-Cho, ALP, and TP (Fig. 1). High discriminative ability for GD: (AUC 0.97, positive predictive value 97.8%, accuracy 92.9%) in the test set was observed even when we

| Number | Graves Disease | Painless thyroiditis | p value |
|--------|----------------|----------------------|---------|
| Men:Female | 3,503:15,832 | 480:2,787 | < 0.001 |
| Age (yrs) | 39.9 (14.1) | 39.5 (14.2) | p = 0.009 |
| RBC (×10^6/μL) | 465.1 (46.0) | 446.1 (40.6) | p < 0.001 |
| Hb (g/dL) | 13.2 (1.3) | 13.2 (1.2) | p = 0.84 |
| Ht (%) | 39.5 (3.7) | 39.4 (3.4) | p = 0.06 |
| MCV (fL) | 85.2 (5.1) | 88.4 (4.9) | p < 0.001 |
| MCH (pg) | 28.5 (2.0) | 29.7 (2.0) | p < 0.001 |
| MCHC (%) | 33.5 (0.9) | 33.5 (0.9) | p < 0.001 |
| Plt (×10^6/μL) | 24.4 (5.8) | 25.4 (6.2) | p < 0.001 |
| WBC (μL) | 5,824.8 (1,680.3) | 5,732 (1,616.5) | p = 0.0504 |
| Neu (%) | 50.6 (10.4) | 56.4 (9.6) | p < 0.001 |
| Lym (%) | 36.3 (9.1) | 31.5 (8.2) | p < 0.001 |
| Mo (%) | 10.0 (3.2) | 9.0 (3.0) | p < 0.001 |
| Eo (%) | 2.6 (2.4) | 2.6 (2.2) | p = 0.2 |
| Ba (%) | 0.6 (0.6) | 0.7 (0.5) | p < 0.001 |
| TP (g/dL) | 6.9 (0.5) | 7.1 (0.5) | p < 0.001 |
| T-Bil (mg/dL) | 0.7 (0.3) | 0.7 (0.3) | p < 0.001 |
| AST (U/L) | 29.1 (15.9) | 24.8 (14.1) | p < 0.001 |
| ALT (U/L) | 39.5 (30.4) | 30.7 (29.7) | p < 0.001 |
| LDH (U/L) | 165.9 (29.9) | 163.6 (30.3) | p < 0.001 |
| γGTP (U/L) | 39.2 (36.2) | 27.3 (31.8) | p < 0.001 |
| ALP (U/L) | 355.2 (175.6) | 218.9 (105.0) | p < 0.001 |
| ChE (U/L) | 387.3 (81.9) | 355.7 (80.3) | p < 0.001 |
| CPK (U/L) | 62.3 (101.1) | 72.7 (156.1) | p < 0.001 |
| CRE (mg/dL) | 0.5 (0.2) | 0.6 (0.1) | p < 0.001 |
| UA (mg/dL) | 5.1 (1.2) | 4.7 (1.1) | p < 0.001 |
| Na (mmol/L) | 140.0 (2.0) | 140.0 (1.9) | p < 0.001 |
| K (mmol/L) | 4.3 (0.3) | 4.3 (0.3) | p = 0.019 |
| Cl (mmol/L) | 105.3 (2.3) | 105.2 (2.2) | p = 0.0012 |
| Ca (mg/dL) | 9.6 (0.4) | 9.5 (0.4) | p < 0.001 |
| P (mg/dL) | 4.0 (0.7) | 3.7 (0.6) | p < 0.001 |
| T-C (mg/dL) | 148.0 (31.3) | 164.5 (34.0) | p < 0.001 |
restricted the predictors in the model to these five variables (age and serum CRE, T-Cho, ALP, and TP) (Fig. 2A), and the prediction model based on logistic regression that included these five factors also showed high discriminative ability for GD (AUC 0.95) (Fig. 2B).

**Differentiation between GD and PT**

The training dataset (70% of the random sample) consisted of the data of 13,544 GD patients and 2288 PT patients, and the test dataset (30% of the random sample) consisted of the data of 5791 GD patients and 979 PT patients. In the test set, our machine learning-based prediction model showed high discriminative ability (AUC 0.89, positive predictive value 93.5%, accuracy 86.0%). The important predictors for the diagnosis of GD included serum ALP, CRE, TP, and γGTP, and WBC. The prediction

**Table 3** The characteristics of the patients with thyrotoxicosis (GD + PT) and normal subjects

| Number | Thyrotoxicosis | Normal subjects | p value |
|--------|----------------|----------------|---------|
| Men:Female | 22,602 | 4,159 | < 0.001 |
| Age (yrs) | 39.8 (14.2) | 34.3 (13.4) | < 0.001 |
| RBC (×10³/µL) | 462.4 (45.8) | 446.8 (42.7) | < 0.001 |
| Hb (g/dL) | 13.2 (1.3) | 13.6 (1.3) | < 0.001 |
| Ht (%) | 39.5 (3.7) | 40.3 (3.8) | < 0.001 |
| MCV (fL) | 85.6 (5.2) | 90.2 (5.0) | < 0.001 |
| MCH (pg) | 28.7 (2.1) | 30.4 (2.1) | < 0.001 |
| MCHC (%) | 33.5 (0.9) | 33.7 (0.8) | < 0.001 |
| Pt (×10³/µL) | 24.6 (5.9) | 24.8 (5.5) | < 0.001 |
| WBC (µL) | 5,811.5 (1,671.5) | 6,124.2 (1,643.4) | < 0.001 |
| Neu (%) | 51.4 (10.5) | 57.0 (9.3) | < 0.001 |
| Lym (%) | 35.6 (9.1) | 32.0 (8.1) | < 0.001 |
| Mo (%) | 9.84 (3.2) | 7.4 (2.1) | < 0.001 |
| Eo (%) | 2.6 (2.4) | 2.8 (2.5) | < 0.001 |
| Ba (%) | 0.6 (0.6) | 1.0 (0.5) | < 0.001 |
| TP (g/dL) | 6.9 (0.5) | 7.3 (0.4) | < 0.001 |
| T-Bil (mg/dL) | 0.7 (0.3) | 0.8 (0.3) | < 0.001 |
| AST (U/L) | 28.5 (15.7) | 20.6 (9.8) | < 0.001 |
| ALT (U/L) | 38.2 (30.5) | 19.5 (15.9) | < 0.001 |
| LDH (U/L) | 165.5 (30.0) | 168.8 (34.0) | < 0.001 |
| γGTP (U/L) | 37.5 (35.8) | 25.0 (48.1) | < 0.001 |
| ALP (U/L) | 335.5 (173.9) | 210.9 (128.5) | < 0.001 |
| ChE (U/L) | 382.7 (82.4) | 303.3 (72.8) | < 0.001 |
| CPK (U/L) | 63.8 (110.8) | 99.4 (171.1) | < 0.001 |
| CRE (mg/dL) | 0.5 (0.2) | 0.7 (0.1) | < 0.001 |
| UA (mg/dL) | 5.0 (1.2) | 4.6 (1.2) | < 0.001 |
| Na (mmol/L) | 140.0 (2.0) | 139.4 (1.9) | < 0.001 |
| K (mmol/L) | 4.3 (0.3) | 4.2 (0.3) | < 0.001 |
| Cl (mmol/L) | 105.3 (2.3) | 104.0 (2.3) | < 0.001 |
| Ca (mg/dL) | 9.6 (0.4) | 9.5 (0.4) | < 0.001 |
| P (mg/dL) | 3.9 (0.7) | 3.6 (0.6) | < 0.001 |
| T-C (mg/dL) | 150.4 (32.2) | 190.5 (35.5) | < 0.001 |
model based on logistic regression also showed high discriminative ability for GD (AUC 0.86) but similar or slightly lower than the machine learning-based model.

**Prediction of thyrotoxicosis**

The training dataset (70% of the random sample) consisted of the data of 15,832 patients with thyrotoxicosis and 2,911 normal subjects, and the test dataset (30% of the random sample) consisted of the data of 6,770 patients with thyrotoxicosis and 1,248 normal subjects. In the test set, our machine learning-based prediction model showed high discriminative ability (AUC 0.98, positive predictive value 98.2%, accuracy 93.7%). The important predictors for the diagnosis of thyrotoxicosis included serum CRE, T-Cho, ChE, and CPK, and Basophils. The prediction model based on logistic regression also showed high discriminative ability for GD (AUC 0.97) but similar or slightly lower than the machine learning-based model.

**Prediction of GD according to sex**

The female training dataset included the data of 11,121 GD patients and 2,339 normal subjects. Our machine learning-based prediction model for the diagnosis of GD was evaluated in the test dataset (30% random sample: 4,711 GD patients and 990 normal subjects). In the test set, our model showed high discriminative ability as demonstrated by the ROC curves; the AUC was 0.99 (Fig. 3A), accuracy 95.5%, and precision 98.73%. The important predictors for the diagnosis of GD among women included CRE, T-Cho, ChE, ALP, and CPK. The logistic regression model showed an AUC of 0.98 was obtained with a prediction model based on logistic regression.

The male training dataset included the data of 2,423 GD patients and 572 normal subjects. In the test set (1,080 GD patients and 258 normal subjects), our model showed high discriminative ability as demonstrated by the ROC curves; the AUC was 0.99 (Fig. 3B, accuracy 94.4%, and precision 98.4%). The important predictors for the diagnosis of GD among men were ALP, CRE, T-Cho, TP, and ALT. An AUC of 0.98 was obtained with the prediction model based on logistic regression.
Prediction of mild and severe thyrotoxicosis due to GD

In the analysis for severe GD, the training dataset included the data of 6,040 patients with severe GD and 2,911 normal subjects. The prediction model for the diagnosis of severe GD was evaluated in the test dataset (30% random sample: 2,516 GD patients and 1,248 normal subjects). Our model showed high discriminative ability as demonstrated by the ROC curves; the AUC was 0.9997 (Fig. 3C), accuracy 99.55%, and precision 99.76%. The strong predictors for the diagnosis of severe GD were T-Cho, CRE, TP, ChE, and Monocytes. A prediction model based on logistic regression also showed high discriminative ability for GD (AUC 0.9993) but lower than the machine learning-based model.

In the analysis for mild GD, the training dataset included the data of 7,505 patients with mild GD and 2,911 normal subjects. The prediction model for the diagnosis of mild GD was evaluated in the test dataset (30% random sample: 3,275 mild-GD patients and 1,248 normal subjects). Our model showed high discriminative ability as demonstrated by the ROC curves; the AUC was 0.98 (Fig. 3D), accuracy 92.8%, and precision 96.38%. The strong predictors for the diagnosis of mild GD were CRE, T-Cho, age, ALP, and TP. The prediction model based on logistic regression also showed high discriminative ability for GD (AUC 0.97) but similar or slightly lower than the machine learning-based model.

Additional analysis (Joint research with a health checkup facility)

We conducted an additional joint study with a health checkup facility. The data of 11,525 subjects who underwent a general health checkup that included thyroid function tests at the Midtown Clinic were included. Age, sex, and the results of the general blood tests alone were used first to identify subjects at risk for thyrotoxicosis by fitting the former prediction model. Since CPK was not present in the checkup, CPK was not included in the analysis.

Fig. 3  (A) Ability of machine learning model to predict GD among women in the test set, and the contribution of each predictor. (B) Ability of machine learning model to predict GD among men in the test set, and the contribution of each predictor. (C) Ability of machine learning model to predict severe GD in the test set, and the contribution of each predictor. (D) Ability of machine learning model to predict mild GD in the test set, and the contribution of each predictor.
measured, CPK was excluded from the prediction model. The subjects’ median age was 51 years old (range: 18 to 91), and there were 6,944 males and 4,581 females. The results showed that the model identified 1,756 suspected cases of thyrotoxicosis among the 11,525 subjects. The actual thyroid function tests revealed thyrotoxicosis in 21 cases (0.2%), 18 of which were suspected of thyrotoxicosis by the predictive model, and the other 3 cases were classified as no suspicion of thyroid disease. The prediction model had an accuracy of 85.7%, sensitivity of 84.9%, specificity of 84.9%, positive predictive value of 1%, and negative predictive value of 99%.

Discussion

Based on a large database (19,335 GD patients, 3,267 PT patients, and 4,159 normal subjects), both our prediction model built with the artificial intelligence software Prediction One and our logistic regression model showed high predictive ability for GD based on the complete blood count and biochemistry parameters, and high predictive performance was achieved even when we included only the five most important variables (age and serum CRE, T-Cho, ALP, and TP) selected with the software. Since our hospital specializes in thyroid disorders, all outpatient clinic visitors had undergone thyroid ultrasonography and blood tests, including measurement of serum thyroid hormone levels and testing for the presence of antithyroid antibodies. Thyrotoxicosis gives rise to a variety of symptoms such as palpitations, tachycardia, weight loss, fatigue, and tremors. Since such symptoms occur in various diseases and are not specific to thyrotoxicosis, it is difficult to make a diagnosis of thyrotoxicosis without measuring thyroid hormone levels [10]. GD is an autoimmune disorder in which thyroid stimulating antibodies stimulate excessive thyroid hormone production [11]. Ophthalmopathy and a palpable goiter are the most common manifestations of GD. A RAIU test is helpful in determining the cause of thyrotoxicosis, but few facilities are capable of performing RAIU tests.

The prediction model using Prediction One enabled us to predict thyrotoxicosis and differentiate between GD and PT with high predictive values based on the complete blood count and biochemistry parameter values. The strong predictors of a diagnosis of GD in our normal subject group were age and serum CRE, T-Cho, ALP, and TP. The strong predictors of diagnosis of GD in the PT group were serum ALP, CRE, TP, and γGTP, and WBC. Serum CRE and ALP were both strong predictors of GD among the normal subjects and the PT subjects. The strong predictors for the diagnosis of thyrotoxicosis in the normal subject group were serum CRE, T-Cho, ChE, and CPK, and basophils. Serum CRE has been found to be a strong predictor of thyrotoxicosis, and in the present study the serum CRE levels were lower in the thyrotoxicosis group than in the normal subject group. It is well known that serum CRE levels are high in hypothyroidism [12, 13]. In thyrotoxicosis, CRE secretion in the renal tubes increases, and the serum CRE level decreases as a result. Muscle mass also decreases in thyrotoxicosis, and that too may lead to a low serum CRE level. Elevated liver enzyme levels are often detected at the onset of GD, and the reported prevalence of elevated liver enzyme levels during the acute thyrotoxic phase of GD has ranged from 37% to 78%. Most studies have described ALT and ALP as the most commonly elevated liver enzymes in the thyrotoxic phase of GD. It is unclear whether bone-derived ALP contributed to the ALP elevations that have been reported in the thyrotoxic phase of GD [14-16]. The duration of thyrotoxicosis in GD patients before the time of diagnosis might be longer than in PT patients, since thyrotoxicosis is transient in PT. The serum cholesterol level is well known to be suppressed in thyrotoxicosis. There is an increase in both cholesterol synthesis and degradation in thyrotoxicosis, but the balance results in a new lower steady-state serum concentration. A low total serum protein level is also a strong predictor of GD, but the mechanism responsible for the decline is unknown. A low WBC count was a strong predictor of GD among the PT patients, and the percentage of basophils was a strong predictor of thyrotoxicosis, but the mechanism responsible for the decreased percentage of basophils is unknown.

According to sex, the important predictors for the diagnosis of GD among women consisted of CRE, T-Cho, ChE, ALP, and CPK, and the important predictors among men consisted of ALP, CRE, T-Cho, TP, and ALT. Serum CRE, ALP, and T-Cho were the strongest predictors in both sexes. The analysis according to the severity of GD showed that the strong predictors for a diagnosis of severe GD were T-Cho, CRE, TP, ChE, and Monocytes, and that the strong predictors for a diagnosis of mild GD were CRE, T-Cho, age, ALP, and TP. The mechanism responsible for the differences in the profiles is unclear, but it is interesting that the profiles differed according to sex and the severity of thyrotoxicosis.

In the additional joint study with a health checkup facility, 1,756 of the 11,525 subjects were suspected of thyrotoxicosis based on the Prediction One prediction model. The actual thyroid function tests revealed thyrotoxicosis in 21 cases (0.2%), 18 of which were suspected of thyrotoxicosis according to the prediction model, and the other 3 cases were classified as no suspicion of thyroid disease. There are differences between prevalences of disease obtained when groups of presumably healthy
people are screened, such as during physical examinations and checkups, and prevalences obtained among groups of people who consult medical institutions because they have symptoms. When prevalences are low, there are many false positives [17]. Even when a test has 99.9% sensitivity and 99.9% specificity, the positive result rate is 91% when the prevalence is 1%, and 50% when the prevalence is 0.1%. Due to the low prevalence of thyrotoxicosis in the group of subjects in the health checkup facility, the positive predictive value was only 1%, but the prediction model was effective in narrowing down the number of suspected cases of thyrotoxicosis to 15% (1,756/11,525).

This study had several limitations. First, because our hospital specializes in thyroid disorders, the subjects who came to our hospital were concerned that they might have a thyroid disorder based on their symptoms, and thus the healthy control subjects may not be representative of all healthy people. Second, our data were cross-sectional, and because the average age of the cohort was less than 40 years and there was a high proportion of women, the cohort was not representative of the entire Japanese population. Since age and gender have an effect on biochemistry profiles and different measurement kits may have different reference values, our prediction models for a diagnosis of GD might not be generalizable to the prediction of incident GD. Third, Prediction One automatically adjusts and optimizes the variables and creates the best prediction model by means of an artificial neural network and gradient boosting decision tree with internal cross-validation, the details are trade secrets. Even the detailed algorithm has not been fully made public, the accuracy of the software seems clear, since its sensitivity and specificity were not inferior to the sensitivity and specificity of the logistic regression model. Fourth, we excluded patients who had been treated for thyroid disorders before their first visit to our hospital and patients who had been taking medication that might affect thyroid function. Patients taking medication for hypertension or hyperlipidemia were not excluded. Fifth, there is a lack of information about symptoms related to thyroid diseases. Lastly, we did not have information in regard to certain risk factors or predictive factors of GD, including smoking and family history of autoimmune thyroid disorders.

In conclusion, the results of this study showed that by using artificial intelligence software and logistic regression models it is possible to predict a diagnosis of GD with high discriminative ability on the basis of only information on age and serum CRE, T-Chol, ALP, and TP levels.

Acknowledgments
We thank all the staff of Ito Hospital and patients for follow up.

Author Disclosure Statement
The authors declare that they have no conflicts of interest to report in regard to this study.

Funding Information
No funding was received for this article.

References
1. Kasagi K, Takahashi N, Inoue G, Honda T, Kawachi Y, et al. (2009) Thyroid function in Japanese adults as assessed by a general health checkup system in relation with thyroid-related antibodies and other clinical parameters. Thyroid 19: 937–944.
2. Smith TJ, Hegedüs L (2016) Graves’ disease. N Engl J Med 375: 1552–1565.
3. Vestergaard P, Mosekilde L (2003) Hyperthyroidism, bone mineral, and fracture risk—a meta-analysis. Thyroid 13: 585–593.
4. Chaker L, van den Berg ME, Niemeijer MN, Franco OH, Dehghan A, et al. (2016) Thyroid function and sudden cardiac death: a prospective population-based cohort study. Circulation 134: 713–722.
5. Sheu JJ, Kang JH, Lin HC (2010) Hyperthyroidism and risk of ischemic stroke in young adults. A 5-year follow-up study. Stroke 41: 961–966.
6. Bartalena L (2018) Graves’ Disease: Complications. In: Feingold KR, Anawalt B, Boyce A, Chrousos G, de Herder WW, et al. (eds) Endotext. MDText.com, Inc. Copyright © 2000–2021, MDText.com, Inc., South Dartmouth, Massachusetts, USA.
7. Wiemken TL, Kelley RR (2020) Machine learning in epidemiology and health outcomes research. Annu Rev Public Health 41: 21–36.
8. Peterson ED (2019) Machine learning, predictive analytics, and clinical practice: can the past inform the present? JAMA 322: 2283–2284.
9. Japan Thyroid Association (2019) Guideline for the treatment of Graves’ disease in Japan 2019. Nankodo, Tokyo, Japan (In Japanese).
10. De Leo S, Lee SY, Braverman LE (2016) Hyperthyroidism. Lancet 388: 906–918.
11. Bartalena L (2013) Diagnosis and management of Graves Disease of Graves’ disease
disease: a global overview. Nat Rev Endocrinol 9: 724–734.

12. Piras C, Pibiri M, Leoni VP, Balsamo A, Tronci L, et al. (2021) Analysis of metabolomics profile in hypothyroid patients before and after thyroid hormone replacement. J Endocrinol Invest 44: 1309–1319.

13. Chen Y, Zhang W, Wang N, Wang Y, Wang C, et al. (2020) Thyroid parameters and kidney disorder in type 2 diabetes: results from the METAL study. J Diabetes Res 2020: 4798947.

14. Hsieh A, Adelstein S, McLennan SV, Williams PF, Chua EL, et al. (2019) Liver enzyme profile and progression in association with thyroid autoimmunity in Graves’ disease. Endocrinol Diabetes Metab 2: e00086.

15. Ogrin C (2008) Increased bone alkaline phosphatase and isolated subcortical bone uptake of technetium-99m hydroxymethylene diphosphonate in the lower extremities in a patient with Graves’ disease: a distinctly unusual variant of Graves’ acropachy. Thyroid 18: 1227–1229.

16. Kubota S, Amino N, Matsumoto Y, Ikeda N, Morita S, et al. (2008) Serial changes in liver function tests in patients with thyrotoxicosis induced by graves’ disease and painless thyroiditis. Thyroid 18: 283–287.

17. Cleary PD, Barry MJ, Mayer KH, Brandt AM, Gostin L, et al. (1987) Compulsory premarital screening for the human immunodeficiency virus: technical and public health considerations. JAMA 258: 1757–1762.