International Journal of Medical Informatics 106 (2017) 32–36

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Research Paper

Impact of predicting health-guidance candidates using massive health check-up data: A data-driven analysis

Daisuke Ichikawa[a,*], Toki Saito[b], Hiroshi Oyama[a,b]

[a] Department of Clinical Information Engineering, Division of Social Medicine, Graduate School of Medicine, The University of Tokyo, Japan
[b] Department of Clinical Information Engineering, Division of Health Services Sciences, School of Public Health, Graduate School of Medicine, The University of Tokyo, Japan

ARTICLE INFO

Keywords:
Health checkup
Health guidance
Machine learning
Prediction
Data-driven

ABSTRACT

Introduction: Starting in 2008, specific health checkups and health guidance to prevent non-communicable diseases have been provided in Japan, which has the highest proportion of elderly citizens in the world. The attendance rate for health guidance appointments is 17.7%, which is far from the national goal of the system (45%). To improve the attendance rate, we present a model for predicting whether an examinee is a candidate for health guidance; this model was based on a machine learning method and a restricted but massive amount of health checkup information.

Materials and methods: Using machine learning methods, we developed the following five prediction models for identifying health-guidance candidates: baseline: this model included sex and age; model 1: this model included variables that can be measured in person + information on whether the examinee was a candidate in the past year; model 2: model 1 + systolic blood pressure + diastolic blood pressure; model 3: model 2 + all health checkup results from the past year; and model 4: model 3 using the training dataset excluding cases with missing data.

Results: The performance levels of the five prediction models (the AUC values of the models for the test dataset) were as follows: 0.592 [95% CI: 0.586–0.596] for the baseline model, 0.855 [95% CI: 0.851–0.858] for model 1, 0.985 [95% CI: 0.984–0.985] for model 2, 0.993 [95% CI: 0.993–0.993] for model 3, and 0.943 [95% CI: 0.941–0.945] for model 4.

Conclusions: We studied five models for identifying health-guidance candidates. The model that used all health checkup results from the past year had the highest predictive power. Application of the prediction model developed in the present study to the selection of health-guidance candidates could reduce the cost of guidance.

1. Introduction

The proportion of people over the age of 60 years will double, from about 11% to 22%, between 2000 and 2050 [1–3]. Degenerative aging processes are the major underlying causes of noncommunicable diseases (NCDs), including ischemic heart disease, stroke, and others [4]. Specific health checkups and health guidance to prevent NCDs began to be provided in Japan, which has the highest proportion of elderly citizens in the world, in 2008 [5]. The specific health checkup, which focuses on Metabolic Syndrome (MetS), identifies MetS patients from among health checkup examinees. After the health checkup, MetS patients are provided with specific health guidance, which mainly involves a lifestyle modification program [6].

The attendance rate for health-guidance appointments is 17.7%, which is far from the national goal of the system (45%) [7]. One of the causes of the low attendance rate is the lag between health checkups and health guidance [8]. The selection criteria, which are consistent with the criteria for MetS, consist of physical (waist circumference, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP)) and blood-related (fasting blood glucose (FBG), hemoglobin A1c (HbA1c), triglyceride (TG), high-density lipoprotein cholesterol (HDL)) variables, as well as the individual's history of medication for lifestyle diseases (hypertension, diabetes, and dyslipidemia). As it takes considerable time to acquire the blood test results, candidates cannot be selected on the day of examination [7].

A system in which candidates for health guidance would be identified without blood test results could reduce the lag and increase the attendance rate. We present a prediction model, using a machine

* Corresponding author at: Department of Clinical Information Engineering, Division of Social Medicine, Graduate School of Medicine, The University of Tokyo 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan.
E-mail address: ichikawa-kyu@umin.ac.jp (D. Ichikawa).

http://dx.doi.org/10.1016/j.ijmedinf.2017.08.002
Received 30 April 2017; Received in revised form 20 July 2017; Accepted 2 August 2017

learning method and a restricted but massive amount of health checkup data, for identifying whether an examinee is a candidate for health guidance.

The application of a predictive model using a machine learning method has achieved successful results in preventive medicine [9–13]. Also, managed health insurance companies in Japan can use health checkup data for the last year for a prediction model because the health checkup rate of insured people is very high (84%) [14].

The purpose of our present study was to develop a prediction model for identifying health-guidance candidates using this massive amount of checkup information and to evaluate the model.

2. Materials and methods

2.1. Health checkup data

We used the nationwide health checkup database obtained from 35 health insurance companies in Japan. The study population consisted of people aged 40–60 years who had no history of medication use and who underwent specific health checkups throughout Japan between April 1, 2014 and March 31, 2016. The medical records of the population were used, and the following health checkup variables were considered: age, sex, BMI (calculated as body weight (kg) divided by the square of height (m^2)), tests for high blood pressure SBP, diastolic (DBP), tests for diabetes (HbA1c, FBG), tests for dyslipidemia (TG, HDL), and history of medication for lifestyle diseases (hypertension, diabetes, dyslipidemia). The blood test results were validated by external quality assessments [15].

We separated the dataset into a training dataset and a test dataset to build the classification model. An 80% randomly sampled dataset of the total dataset was used as the training dataset, and the remaining 20% was used as the test dataset. The final training and test datasets were composed of 224,130 and 56,033 persons, respectively. Table 1 lists the demographic and clinical characteristics of the subjects.

The health-guidance candidates were judged based on the specific health guidance criteria in Japan [6]. They were health checkup examinees who were obese and had cardiovascular risks. Specifically, the criteria have two parts: one assesses obesity risk (waist circumference or BMI), and the other assesses cardiovascular risk (SBP, DBP, FBG, HbA1c, TG, and HDL). The examinees were identified as health-guidance candidates when they fulfilled both criteria.

The numbers of health-guidance candidates in the training and test datasets were 60,043 (26.8% of the total) and 14,867 (26.5% of the total), respectively. The proportions of health-guidance candidates who had obesity risks were 78.6% in the training dataset and 78.4% in the test dataset.

The protocol of the present study was approved by the Review Board of Life Science Research Ethics and Safety of the University of Tokyo (#15-137).

Table 2
Numbers of records with missing data for FBG and HbA1c in the training and test datasets in 2014 and 2015.

Variable	Training dataset		Test dataset	
	2014	2015	2014	2015
FBG	13.4%	13.3%	13.5%	13.3%
HbA1c	18.0%	17.4%	18.1%	17.5%

2.2. Data processing

We standardized the training and test datasets by transforming each continuous variable so that it had zero mean and unit variance by subtracting the mean of each continuous variable divided by the standard deviation.

2.3. Missing data

All results used in the selection of candidates for health guidance except FBG and HbA1c were mandatory parts of the health checkup. Either FBG or HbA1c is required in the health checkup. Therefore, except with regard to FBG and HbA1c, there were no missing data in any of the datasets used in the present study. The numbers of records with missing data for FBG and HbA1c in the training and test datasets in both 2014 and 2015 are described in Table 2. Because the amount of missing data could not be ignored, we used knnImpute as the input method; this is an input method that utilizes the k-nearest neighbor method [16]. We set k = 5 for knnImpute. The method was implemented using “R” software (version 3.3.2) and the R package “caret” [17,18]. Furthermore, we compared the prediction model using excluding cases with missing data.

2.4. Prediction model

We compared the five prediction models for health-guidance candidates: baseline: a model with sex and age as variables; model 1: a model with variables that can be measured in person + whether the examinee was a candidate in the past year; model 2: model 1 + SBP + DBP; model 3: model 2 + all health checkup results from the past year; model 4: model 3 using the training dataset excluding cases with missing data (Table 3).

Other than the baseline model, we used a gradient-boosting decision tree (GBDT) [19] as the machine learning method. Ensemble learning refers to an algorithm that combines base learners (such as decision trees and linear classifiers); the GBDT approaches are examples of such

Table 1
Characteristics of datasets.

Variable	Training dataset (n = 224,130)		Test dataset (n = 56,033)	
	Former (year 2014) mean (SD)	Latter (year 2015) mean (SD)	Former (year 2014) mean (SD)	Latter (year 2015) mean (SD)
Age		50.7 (5.1)		50.7 (5.1)
Sex (Female)		20.1 (45,058)		20 (11,227)
# of Candidates	25.9 (58,069)	26.8 (60,043)	25.6 (14,366)	26.5 (14,867)
BMI	22.9 (3.2)	22.9 (3.2)	22.9 (3.1)	22.9 (3.2)
Waist circumference	81.5 (8.7)	81.6 (8.8)	81.5 (8.7)	81.6 (8.7)
Systolic blood pressure	119.5 (15)	120.2 (15.5)	119.6 (15)	120.3 (15.5)
Diastolic blood pressure	75.6 (11.1)	76.1 (11.4)	75.7 (11.1)	76.1 (11.4)
Fasting blood glucose	94.6 (12.6)	94.9 (13.4)	94.5 (12.3)	94.8 (13.3)
HbA1c	5.5 (0.4)	5.5 (3.9)	5.5 (0.4)	5.5 (0.4)
Triglyceride	111.7 (89.2)	111.5 (89.9)	111.7 (87.3)	111.1 (89.3)
HDL-cholesterol	62.7 (16.6)	62.8 (16.7)	62.6 (16.5)	62.8 (16.7)