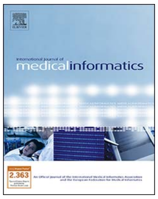
Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



Prediction models to identify individuals at risk of metabolic syndrome who are unlikely to participate in a health intervention program

Akihiro Shimoda[a,*], Daisuke Ichikawa[a], Hiroshi Oyama[a,b]

[a] *Department of Clinical Information Engineering, Division of Social Medicine, Graduate School of Medicine, the University of Tokyo, Bunkyo-ku, Tokyo, Japan*
[b] *Department of Clinical Information Engineering, School of Public Health, Graduate School of Medicine, the University of Tokyo, Bunkyo-ku, Tokyo, Japan*

ARTICLE INFO

Keywords:
Health check-up
Instruction program
Machine learning
Prediction

ABSTRACT

Objectives: Since the launch of a nationwide general health check-up and instruction program in Japan in 2008, interest in strategies to improve implementation of the program based on predictive analytics has grown. We investigated the performance of prediction models developed to identify individuals classified as “requiring instruction” (high-risk) who were unlikely to participate in a health intervention program.

Methods: Data were obtained from one large health insurance union in Japan. The study population included individuals who underwent at least one general health check-up between 2008 and 2013 and were identified as “requiring instruction” in 2013. We developed three prediction models based on the gradient boosted trees (GBT), random forest (RF), and logistic regression (LR) algorithms using machine-learning techniques and compared the areas under the curve (AUC) of the developed models with those of two conventional methods The aim of the models was to identify at-risk individuals who were unlikely to participate in the instruction program in 2013 after being classified as requiring instruction at their general health check-up that year.

Results: At first we performed the analysis using data without multiple imputation. The AUC values for the GBT, RF, and LR prediction models and conventional methods: 1, and 2 were 0.893 (95%CI: 0.882–0.905), 0.889 (95%CI: 0.877–0.901), 0.885 (95%CI: 0.872–0.897), 0.784 (95%CI: 0.767–0.800), and 0.757 (95%CI: 0.741–0.773), respectively. Subsequently, we performed the analysis using data after multiple imputation. The AUC values for the GBT, RF, and LR prediction models and conventional methods: 1, and 2 were 0.894 (95%CI: 0.882–0.906), 0.889 (95%CI: 0.887–0.901), 0.885 (95%CI: 0.872–0.898), 0.784 (95%CI: 0.767–0.800), and 0.757 (95%CI: 0.741–0.773), respectively. In both analyses, the GBT model showed the highest AUC among that of other models, and statistically significant difference were found in comparison with the LR model, conventional method 1, and conventional method 2.

Conclusion: The prediction models using machine-learning techniques outperformed existing conventional methods: for predicting participation in the instruction program among participants identified as “requiring instruction” (high-risk).

1. Introduction

The burden of non-communicable disease (NCD) has increased in recent decades due to historic and projected demographic shifts in the population, urbanization, and dramatic lifestyle changes worldwide [1]. In Japan, NCDs, including heart disease, cancer, stroke, and diabetes are the leading causes of death, and 79% of total deaths were estimated to be derived from NCDs in 2012 [2]. The major determinants of NCD incidence and mortality are behavioral risk factors including smoking tobacco, drinking alcohol, physical inactivity, sedentary behavior, and obesity [3,4]. Thus, promoting a healthier lifestyle in the general population has become a focus of public health practitioners [5].

Several countries provide nationwide health check-up programs to promote lifestyle modification and improve health in the general public [6–8]. Japan initiated a specific nationwide health check-up and instruction system in April 2008 [9]. The aim of the program was to detect metabolic syndrome in participants and provide individual instruction or necessary treatment to improve the lifestyle of those at high-risk of developing the condition. However, despite considerable promotion, the system has not been effective: although the participation rate for the health check-up was about 50% in 2014, only 17.8% of

[*] 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: shimoda-tky@umin.ac.jp (A. Shimoda).

https://doi.org/10.1016/j.ijmedinf.2017.12.009
Received 25 March 2017; Received in revised form 11 December 2017; Accepted 14 December 2017

the individuals identified as high-risk participated in the program to modify lifestyle [10]. Thus, there has been growing interest in strategies to improve participation in the instruction program among individuals in the high-risk group.

When individuals identified as having or being at high risk of metabolic syndrome at the general health check-up do not immediately take up the instruction program, the healthcare provider is required to intervene to increase the participation rate. At present, letters, emails, or other forms of communication are sent to all non-participants within a specified timeframe; however, it would be more efficient and cost-effective to contact only those non-participants who are highly unlikely to undertake the program. Predictive analytics based on electronic health records (EHRs) held by medical insurers (service providers of specific health check-up and instruction programs) may play an important role in solving this problem. Previous studies have concluded that the application of big data to health care is inevitable [11] and is occurring in Japan. The Japanese Ministry of Health, Labor and Welfare launched the concept of “data health”, set up by medical insurers, and supports the promotion of effective and efficient health programs developed using EHRs [12]. In fact, since the launch, all insurers have developed various interventions incorporating the concept of “data health”. Nevertheless, evidence concerning the effectiveness of preventative health promotion based on predictive analytics is sparse. We believe predictive models that can identify individuals who are unlikely to participate in the instruction program will increase appropriate selection of the target population and, thus, improve the efficacy and efficiency of the program.

The purpose of our study is to examine the effectiveness of predictive analytics to investigate methods to target high-risk individuals who were unlikely to participate in intervention programs. Specifically, we used machine-learning techniques to develop prediction models.

2. Methods

2.1. Study design

Our study is a retrospective analysis of individuals identified as requiring instruction at their general health check-up for identifying those who are unlikely to participate in the instruction program. The study was approved by the Institutional Review Board of the University of Tokyo (Examination number: 10831-[1]) and adapted to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement for reporting prediction models [13].

2.2. Study population and setting

The data were obtained from a large health insurance union in Japan. The study population consisted of individuals aged 35–75 years who had undergone at least one general health check-up between 2008 and 2013 and were identified as “requiring instruction” (high-risk) in 2013. The definition of “requiring instruction” is determined by Japanese criteria [14].

2.3. Study protocol

2.3.1. Dataset creation and definitions

We used only data available or generated during general health check-ups as prediction variables. Relevant data were extracted from the health information system of the Health Insurance Union including demographic information (sex, age, weight, height, BMI, and abdominal circumference), examination results (systolic blood pressure, diastolic blood pressure, triglyceride, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, glutamate oxaloacetate transaminase (AST), ɤ-glutamyl transpeptidase, glycated hemoglobin, and diagnosis), questionnaire responses (anemia, smoking history, weight gain, physical activity, walking speed, eating habits, drinking, sleeping, intention to improve lifestyle, and intention to undergo instruction), and past experience of participation in general health check-ups and instruction. All descriptive statistics were analyzed using R (version 3.3.2) [15]. The list of all predictor variables used are represented in Table 1, and the data processing steps are shown in Fig. 1.

Table 1
List of all predictor variable used for developing models.

Variable type	Variable
Demographics	Sex, Age, Height, Weight, BMI, Abdominal circumference
Examination results	Systolic blood pressure, Diastolic blood pressure, Tryglyceride, High-density lipoprotein cholesterol, Low-density lipoprotein cholesterol, Glutamate oxaloacetate transaminase, ɤ-glutamyl transpeptidase, Glycated hemoglobin, Diagnosis in general health check-up
Questionnaire results	Anemia, Smoking, Weight gain, Exercise, Physical activity, Walking speed, Weight change, Eating habit, Drinking, Sleeping, Intention to improve lifestyle, Intention to undergo instruction
Past participation	Past participation in general health check-up (2008–2012), Past participation in instruction program (2008–2012)

2.3.2. Data preprocessing

All continuous variables were divided into two to five discrete categories with consideration given to the balance between interpretability and reduction in cluster variance (the coordinate-wise squared deviations from the mean of the cluster of all the observations belonging to that cluster). We assumed that missing data were random, depending on the clinical variables measured and the questionnaire responses sought in the medical institutions. The number of records with missing data for systolic blood pressure, diastolic blood pressure, glycated hemoglobin, drinking habit were 2295 (15.8%), 2295 (15.8%), 1047 (7.2%), and 697 (4.8%), respectively. Since the amount of missing data could not be ignored, we imputed missing values using the multiple imputation (MI) program [16,17], to make full use of the data and improve the statistical power. MI has emerged recently as an excellent method for handling missing data because of its flexibility and robustness. The MI uses a multiple simulated version approach. In other words, the MI program replaces missing values with several different

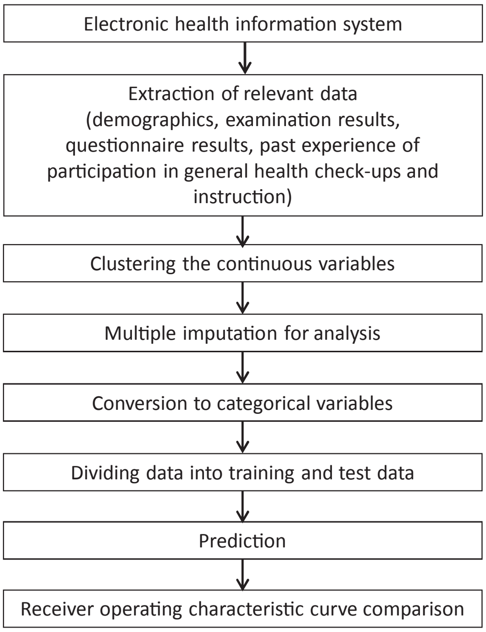

Electronic health information system
Extraction of relevant data (demographics, examination results, questionnaire results, past experience of participation in general health check-ups and instruction)
Clustering the continuous variables
Multiple imputation for analysis
Conversion to categorical variables
Dividing data into training and test data
Prediction
Receiver operating characteristic curve comparison

Fig. 1. Steps for data processing and predictive analyses of this study.