2023; 27: 8829-8841

# Predicting habitual water intake from lifestyle questions

H. TANAKA[1], K. MIZUMA[2], Y. NAKAMURA[3], A. HIRATA[4], J. MIYAZAKI[4], K. SUZUKI[2], H. SETA[2], H. WATANABE[3], T. SUZUKI[3], R. WATANABE[2], N. MURAYAMA[3], T. OKAMURA[4], S. NAKAMURA[1]

[1]Division of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan
[2]Suntory Global Innovation Center Limited, Institute for Water Science, Seika-cho, Soraku-gun, Kyoto, Japan
[3]Suntory Global Innovation Center Limited, Research Institute, Seika-cho, Soraku-gun, Kyoto, Japan
[4]Department of Preventive Medicine and Public Health, Keio University School of Medicine, Shinjuku-ku, Tokyo, Japan

**Abstract. – OBJECTIVE: Previous studies have used selective recall and descriptive dietary record methods, requiring considerable effort for assessing food and water intake. This study created a simplified lifestyle questionnaire to predict habitual water intake (SQW), accurately and quickly assessing the habitual water intake. We also evaluated the validity using descriptive dietary records as a cross-sectional study.**

**SUBJECTS AND METHODS: First, we used crowdsourcing and machine learning to collect data, predict water intake records, and create questionnaires. We collected 305 lifestyle-related questions as predictor variables and selective recall methods for assessing water intake as an outcome variable. Random forests were used for the machine learning models because of their interpretability and accurate estimation. Random forest and single regression correlation analysis were augmented by the synthetic minority oversampling that trained the model. We separated the data by sex and evaluated our model using unseen hold-out testing data, predicting the individual and overall habitual water intake from various sources, including non-alcoholic beverages, alcohol, and food.**

**RESULTS: We found a 0.60 Spearman's correlation coefficient for total water intake between the predicted and the selective recall method values, reflecting the target value to be achieved. This question set was then used for feasibility tests. The descriptive dietary record method helped to obtain a ground-truth value. We categorized the data by gender, season, and source: non-alcoholic beverages, alcohol, food, and total water intake, and the correlation was confirmed. Consequently, our results showed a Pearson's correlation coefficient of 0.50 for total water intake between the predicted and the selective recall method values.**

**CONCLUSIONS: We hypothesize that dissemination of SQW can lead to better health management by easily determining the habitual water intake.**

*Key Words:*

Habitual water intake, Random forests, Questionnaire, Lifestyle

## Introduction

Water, a vital life component, constitutes 60% of the human body[1,2]. Water critically maintains homeostasis and is a medium for delivering oxygen, nutrients, hormones, and other substances throughout the body, and helps remove waste products and excess substances[3]. Moreover, sufficient water intake is critical for human health, creating an association between water intake and health status. Inadequate water intake increases the risk of renal and cardiovascular diseases, contributes to metabolic disorders[4], and is associated with the onset of cerebral infarction[5]. In addition, studies from physical and cognitive perspectives highlight the association between dehydration and many health issues, like loss of attention, concentration, cognitive, mood, and motor functions, fatigue, and headaches[6,7]. For example, Secher and Ritz[8] established a clear link between dehydration and reduced cognitive performance. Other studies[7-10] showed that drinking more water may improve performance

*Corresponding Author:* Hiroki Tanaka, Ph.D; e-mail: hiroki-tan@is.naist.jp

in an attention test. Therefore, consistent and adequate daily water intake could reduce the disease risk and maintain mental health. Successful prediction of habitual water intake might enable building a conversational agent that recommends drinking water based on monitoring the current water intake[11,12].

However, measuring water intake requires a substantial effort, which includes the descriptive dietary record method and food frequency questionnaires obtained *via* interviews with many questions. Moreover, most of these studies have focused on the limited aspects of behavior and lifestyle. For example, a self-administered food frequency questionnaire was analyzed and validated for assessing food[13,14] and specific amino acid intake[15]. Creating new questionnaires is expensive and time-consuming, requiring the evaluation of validation and reliability. Recent advances in data-driven approaches have enabled dealing with large amounts of data. In addition, crowdsourcing is helpful for efficiently collecting data from human participants to reduce the number of questions[16,17]. Partial least squares (PLS) regression and factor analysis were used in a previous study[18] to predict age and body mass index and to observe question weights. Another previous study[19] used a multi-layer perceptron to predict seasonal water consumption.

Our present study aimed to create a new questionnaire using a simplified lifestyle-based one to predict habitual water intake (SQW) with only ten questions and to evaluate the validity by exploring the relationship between SQW values and descriptive dietary records[20,21]. We first constructed ensemble trees to predict water intake records to create a new SQW. Then, our SQW was used for model validation with different regions, seasons, and populations, including those participating in epidemiological studies and those recruited for this study, and descriptive dietary records were obtained as a ground-truth value. The following sections describe the model construction, results, and model validation.

## Subjects and Methods

### Constructing Models for Predicting Water Intake

#### Data collection

We recruited participants three times using CrowdWork (https://crowdworks.jp) to obtain data from sufficient participants, the primary data collection period was from January 14 to 25, 2020. CrowdWork is one of the largest crowdsourcing platforms in Japan, and a previous study[18] obtained data from the platforms. More than eight hundred thousand people are registered in CrowdWorks, and 797 participants applied for our data collection and were used for the cross-sectional analysis. Eligibility criteria included men and women between the ages of 18 and 85, who were members of CrowdWorks, and those who understood and agreed to participate in this study. We first obtained informed consent from all participants. Over one week, the participants answered questions about their dietary recall. We have attached the corresponding images to represent food and drink amounts that were easy to understand in questionnaires. The selective dietary recall questionnaires required approximately 30 min to complete. The crowd workers were allowed to contact the first author if they faced difficulties. Our questionnaire included two dummy questions regarding sex (male or female) and experience of acupuncture or moxibustion (yes or no). Due to the challenging tasks requiring significant effort to answer all questions, we eliminated participants who failed to concentrate on this by observing the consistency of answers to the two dummy questions (Criteria 3).

#### Lifestyle questions collection

We prepared 305 multidimensional questions in Japanese as predictor variables. The questions were either already validated or manually created and relevant to water intake habits (Figure 1a). The questions were carefully selected by consulting a dietitian. It also included a food frequency questionnaire (FFQg) obtained with permission from a previous study[22] since the frequency of food intake may be affected. Other questions were from the simplified nutritional appetite questionnaire (SNAQ-J)[23] and Pittsburgh Sleep Assessment[24], which calibrates sleep quality and smoking behaviors. We also obtained questionnaires from the National Health and Nutrition Survey[25] and the Japanese version of the Constitution in the Chinese Medicine Questionnaire (CCMQ)[26]. We also adopted the Ten Item Personality Inventory (TIPI-J)[27,28]. We defined a category, a physical constitution based on the CCMQ, as an individual's body condition formed by interactions between genetic and environmental factors. In addition to the validated questionnaires, we

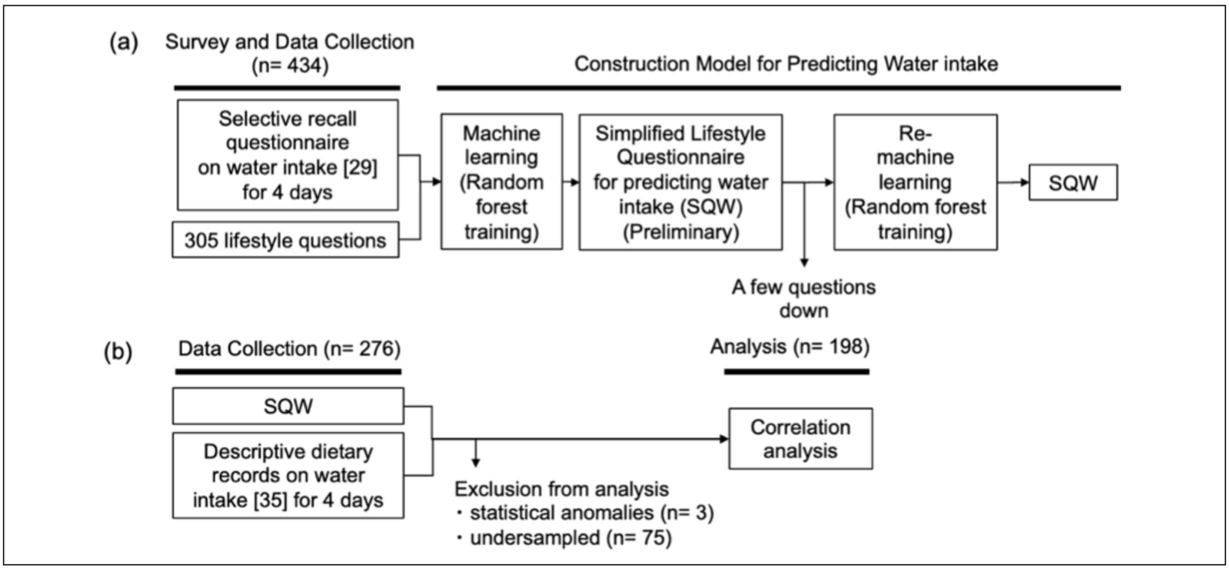


**Figure 1.** Flow diagram of this study. Development of a simplified lifestyle questionnaire for predicting water intake (SQW) (**a**) and validation flow of SQW (**b**).

created 73 new questions regarding dietary and drinking behaviors. A complete list of the 305 questionnaires is available upon request.

### *Selective recall method*

The selective recall method is a survey-style questionnaire for assessing water intake from food and beverage, in which respondents answer the type and amount of food and beverages consumed at each time of the day. Our previous study[29] showed that the correlation coefficient between descriptive dietary recall as a gold standard and selective recall methods was greater than 0.90. To construct the model, we collected daily water intake data based on the selective recall of 215 questions. After the lifestyle questionnaires, the participants addressed the questions on water intake, alcohol use, and food consumption the day before the questionnaire. Since one-day answers were highly biased, they completed a four-day dietary record, which included three weekdays and one weekend day, and used it as an outcome variable. The participants were questioned about their previous day's water and food intake in the morning, afternoon, evening, and night. For example, we collected data on daily water intake from tea, water, coffee, milk, and soft drinks in nine time zones in one day. We then transformed the recalled dietary information into water intake values based on the predicted amount of water in each food and drink. Habitual water intake was averaged over four days, as in the previous study[30].

### *Preprocessing*

We recorded the participants' answers using Google Forms, which exported the gathered data to a CSV text format and processed this in Python (version 3) using the Numpy library[31]. Some features were transformed into binary variables because some answers could not handle the nominal scale of the variables. We obtained 797 participant applications from crowdsourcing and chose 434 individuals (184 males and 250 females) based on the following three selection criteria (Figure 1a). We accepted 56% of the data and rejected 44% of them.

Criteria 1: We removed users who answered more than twice because the Google form accepts multiple answers, which could introduce noise.

Criteria 2: Exact matches of crowdsourcing user IDs that occurred five times: one questionnaire and four diet records.

Criteria 3: We removed participants who gave different answers to the two dummy questions regarding gender and experienced acupuncture or moxibustion.

### *Random forests training*

Before our main study, we attempted several models, including regression trees, PLS regression, and random forests. We used the R packages[32] for statistical and correlation analyses and machine learning algorithms: regression trees, PLS regression, and random forests. One advantage of regression trees is that they use every observation and produce consistent results among trials, enabling an efficient selection of questions. However, regression trees had lower predictive capability than complicated models, including PLS and random forests. Predictive capability is our primary outcome in selecting models for accurately predicting the water intake. We also attempted PLS regression, which generally produces better predictive capability than regression trees. However, the original idea of PLS included finding a good axis representing multiple variables for predicting target values. PLS is not an appropriate prediction model because it uses only a few fixed questions. Random forest performed the best with high interpretability among the above models. The basic concept of the random forest model involves using an ensemble of regression trees. The algorithm has four steps: (1) it draws a random bootstrap sample of size $n$, (2) grows a regression tree from the bootstrap sample (at each node: randomly select $d$ features without replacements, splits the node using the feature for the best split, based on an objective function, such as maximizing the information gain), (3) repeats these steps $m$ times, (4) aggregates the prediction by each tree to assign the class label by averaging the outcomes.

We separated the data as follows: male training, 110; development, 29; test, 45; female training, 150; development, 40; and test, 60. This separation was performed based on random sampling from all data. We ran multiple random forest constructions because they rely on random sampling to validate the stability of the selected features. We attempted numerous separation combinations that showed no significant differences. Our plans to use this trained model in future validation studies led us to fix the training data to be tested. We performed mean age interpolation to avoid unavailable values for training the random forest because of incorrect birth dates typed by some participants. Because we confirmed that water intake is not uniformly distributed (the water intake was imbalanced), we adopted SMOTER[33], an extension of the synthetic minority oversampling technique (SMOTE), for the regression of our models. As shown in Figure 2, the total water intake was imbalanced, especially for outliers' males. Testing our prediction model without such data augmentation led to worse performance than the augmentation. This paper presents the results obtained using SMOTER. We used the following parameters: an indicator of the number of nearest neighbors (=3) and a number indicating the relevance threshold above which a case was considered to belong to the rare class (=0.50). We added them back to our original data to double the number of samples in the training set over the original size. We also performed feature selection based on each of their weight values as a result of the random forest fitting on all the training data. We retained the order of important features (questions) and ran the training, development, and testing again to confirm whether the reduced questions could predict water intake. We attempted to observe the estimation performance between the top 1 and top 15 features because we considered more than 15 questions, which was a substantial number. A random forest has few tuning parameters[34]; therefore, we used a grid search approach to find the appropriate hyperparameters. Based on the grid search, we also determined a grid search for a development dataset with the following values: number of trees: $m$ {100,200,300,400,500}; several features in each tree: $dpost$ ({2,3,4,5,10,15} for post-feature selection), and $dpre$ ({10,20,30,40,50,60,70} for pre-feature selection).

### *Evaluation metrics*

We evaluated our models using testing data. First, we built a model of the training data to determine feature weights and used the selected top $m$ features for retraining. The final evaluation was performed on the test dataset: 45 samples

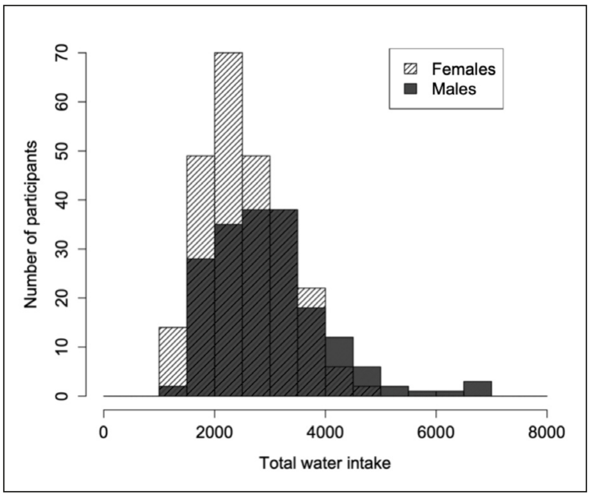


**Figure 2.** Histograms of total water intake (mL) of males and females in crowdsourcing. Water intake is a 4-day average obtained using a selective recall questionnaire on water intake.

from the males and 60 from the females. We also confirmed the highly weighted features produced by random forests. To clarify this ranking, we then interpreted our model's predictions as reasonable and the chosen combination of features as appropriate to reduce the number of questions to 15, which could be answered quickly.

## Ethics Approval

This study was conducted following the guidelines of the Helsinki Declaration (revised by the Fortaleza General Meeting of the World Medical Association). All procedures involving human participants were approved by the Ethics Committee of Suntory Holdings Limited (IRB No. 15000139) and Nara Institute of Science and Technology (IRB No. 2019-I-19). This study also followed the Ethical Guidelines for Medical Research Involving Human Subjects (2014 Ministry of Education, Culture, Sports, Science and Technology and the Ministry of Health, Labour and Welfare Ministerial Notification No. 3). Written informed consent was obtained from all the participants.

## Model Validation

### Data collection

We validated each random forest model using data collected from the Kansai area in Japan between January 2020 and November 2021. For the model validation, we involved different inclusion criteria. Eligibility criteria included men and women over 50 and those who understood and agreed to participate in this study. The recruited participants included 100 males [mean: 58.5, (standard deviation; SD=5.6) years] and 176 females [mean age: 58.1 (SD=4.8) years]. The breakdown is as follows: The KOBE study[35] included 37 male and 121 female subjects. The other, 63 male and 55 female, subjects were from the Oneness Support Co., Ltd.'s panel of subjects. However, there is a possibility that the subjects of this study do not represent the general Japanese population as a selection bias. We collected SQW together with the acquisition of descriptive dietary records for the cross-sectional analysis. The study size was designed with reference to a similar case study[15,36] conducted in the past on the validation of a questionnaire on dietary intake. The model was validated throughout the study period. We assumed that temperature might be a factor influencing water intake, and hence, categorized the three seasons as follows: winter (December, January, February, and March as low-temperature seasons), spring and autumn (April, May, October, and November as mid-temperature seasons), and summer (June, July, August, and September as high-temperature seasons). We eliminated outlier participants based on dietary recall values by using repeated Smirnov-Grubbs tests until no more participants were at $p<0.05$ after confirming the normality of the distribution. For outlier elimination, we did not include the case where water intake was from alcohol because a large amount of data had a value of 0. Overall, three male participants were excluded from the validation analysis. We performed under-sampling of the data in the winter season using random sampling because the winter season has a more significant number of data samples than the other seasons.

### Descriptive dietary records

A descriptive dietary record is a method of writing down all food and drinks, and water intake is calculated from the contents[21]. We collected data on a web form or physically printed paper, depending on the participant's preference. We used the descriptive dietary record method used in a previous study[21,29] as the question we used was the ground truth, which most accurately measured water intake. However, some informational bias in recalling water intake was likely present.

### Validation analysis

The outcome of this study is represented by Pearson's correlations of descriptive dietary records and SQW. The amount of water intake by descriptive dietary record method was analyzed as the dependent variable and SQW as the independent variable. Each data set was analyzed considering the factor of sex.

We validated the constructed models for 276 participants interviewed by dietitians (Figure 1b), following the previously reported ethical guidelines.

## Statistical Analysis

Using the selective recall method, we calculated Spearman's correlation coefficient between the predicted and water intake values. We tested their correlation coefficient to find no correlation with an alpha level of 0.05. Pearson's correlations were calculated between descriptive dietary records and SQW. We used the R packages[32] (The R Foundation for Statistical Computing, Vienna, Austria) for statistical and correlation analyses.

## Results

### Participant's Characteristics

Preprocessing determined the ages of our participants (males, mean: 40.4, SD=11.1; females, mean: 37.3, SD=8.9), with ages ranging from 18 to 73. Figure 2 shows a histogram of the total water intake by all our participants, most of whom drank over 400 mL and less than 2,000 mL. Males drank more than females and showed more physiological outliers, such as those drinking more than 10,000 mL. Alcohol and food-based water intakes are presented in Table I. Alcohol consumption showed high deviations among individuals as the SD was nearly 2-3 times the mean. Moreover, food also contributed to the total water intake, approximately half the amount.

### Correlation Analysis

Table II presents the Spearman's correlation of the water intake values by the selective recall method and predicted values. Based on the values of total, beverage, alcohol, and food intake, we separated the results, which changed by approximately 0.05 based on each trial of the random forest. We plotted the model results and calculated Spearman's correlation coefficient (Figure 3), representing our predictions in all the cases. The results showed a significantly higher correlation than no correlation ($p<0.05$). Our predictions were accurate for people consuming a large amount (above 3,500 mL) and in the mid-range (approximately 2,500 mL). In almost all cases, the correlation coefficients exceeded 0.6, showing a high correlation. In this study, we predicted the habitual water intake from water-, alcohol-, and food-based sources.

Also, Table II shows that the first one of three questions was unstable, but ten questions were stable, with similar values for all 305 inputs. Water intake from food value was relatively difficult to predict at 0.50 for males and 0.37 for females, which confirmed that water intake from food predictions could be reached at all values using up to 30-50 questions.

We performed the same evaluation for beverage-based, alcohol-based, and food-based total water intake, which revealed similar weighted features. However, some of these features were different depending on the source. Other derivations are listed by the following top questions in females:

Water intake from beverage: How many times do you drink tea daily? How much do you usually drink something (mL) in the afternoon?

Water intake from alcohol: The number of alcoholic drinks per week; how often do you drink mineral water instead of tap water?

Water intake from food: How often do you eat vegetables and mushrooms during breakfast? How many cups of miso soup do you drink per week? How many eggs do you eat per week?

### Rechecks of Understandability and Ambiguity of Questions

Before conducting model validation, we further considered whether the SQW question set applies to the general population in terms of understandability and ambiguity. We also reduced the number of questions because some questions had ambiguous answers. We eliminated questions regarding (1) life events (2) whether the participant lived in the Kanto area, and (3) whether the participant was married, as such questions were unrelated to eating, drinking, and life habits. We tested cases in which these questions were removed from the crowdsourced data. The correlation coefficients, excluding the questions, are summarized in Table III. We confirmed the absence of any large difference in the correlations of total water intake by excluding these questions compared with the correlations reported in Table II. Furthermore, the value of water intake of each source was also subjected to re-machine learning with top 9 or top 10 by deleting irrelevant question items, as was the case with total water intake. The correlation coefficient was 0.62 for top 10, 0.78 for top 10, and 0.51 for top 9 for age-derived, alcohol-derived, and food-derived males, respectively, and

**Table I.** Summary of water intake with mean and standard deviation. Water intake is a 4-day average obtained using a selective recall questionnaire on water intake.

| | Total | Beverage | Alcohol | Food |
|---|---|---|---|---|
| Males | 2,998.5 (1,017.7) | 1,658.3 (773.3) | 123.2 (221.7) | 1,217.0 (479.1) |
| Females | 2,566.1 (735.7) | 1,439.8 (578.5) | 56.5 (141.1) | 1,069.7 (319.2) |

**Table II.** All data and top-rank features were measured using Spearman's correlation: the bold value represents the best correlation after feature reduction with minimal features.

| | All | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 6 | Top 7 | Top 8 | Top 9 | Top 10 | Top 11 | Top 12 | Top 13 | Top 14 | Top 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Total** | | | | | | | | | | | | | | | | |
| Males | 0.68 | 0.14 | 0.47 | 0.51 | 0.47 | 0.43 | 0.47 | 0.52 | 0.54 | 0.57 | 0.63 | **0.66** | 0.61 | 0.66 | 0.64 | 0.62 |
| Females | 0.63 | 0.32 | 0.30 | 0.48 | 0.39 | 0.48 | 0.54 | 0.53 | 0.60 | 0.63 | 0.53 | 0.57 | 0.50 | 0.52 | 0.54 | 0.46 |
| **Beverage** | | | | | | | | | | | | | | | | |
| Males | 0.70 | 0.38 | 0.05 | 0.58 | 0.55 | 0.55 | 0.46 | 0.44 | 0.54 | 0.57 | 0.62 | 0.57 | 0.61 | **0.66** | 0.66 | 0.63 |
| Females | 0.68 | -0.22 | 0.04 | 0.45 | 0.61 | 0.55 | 0.61 | 0.56 | 0.59 | 0.53 | 0.57 | 0.60 | 0.61 | 0.60 | 0.65 | 0.61 |
| **Alcohol** | | | | | | | | | | | | | | | | |
| Male | 0.80 | 0.03 | 0.73 | 0.73 | 0.73 | 0.71 | 0.75 | 0.76 | 0.75 | 0.78 | 0.78 | **0.79** | 0.78 | 0.77 | 0.76 | 0.77 |
| Females | 0.73 | -0.16 | 0.70 | 0.71 | 0.64 | 0.65 | 0.67 | **0.72** | 0.69 | 0.70 | 0.70 | 0.68 | 0.66 | 0.67 | 0.63 | 0.67 |
| **Food** | | | | | | | | | | | | | | | | |
| Males | 0.68 | 0.29 | 0.24 | 0.30 | 0.22 | 0.35 | 0.43 | 0.45 | **0.37** | 0.42 | 0.50 | 0.48 | **0.52** | 0.50 | 0.49 | 0.5 |
| Females | 0.47 | 0.26 | 0.11 | 0.18 | 0.12 | 0.21 | 0.31 | 0.33 | 0.32 | 0.37 | 0.33 | 0.32 | 0.29 | 0.33 | 0.35 | 0.31 |

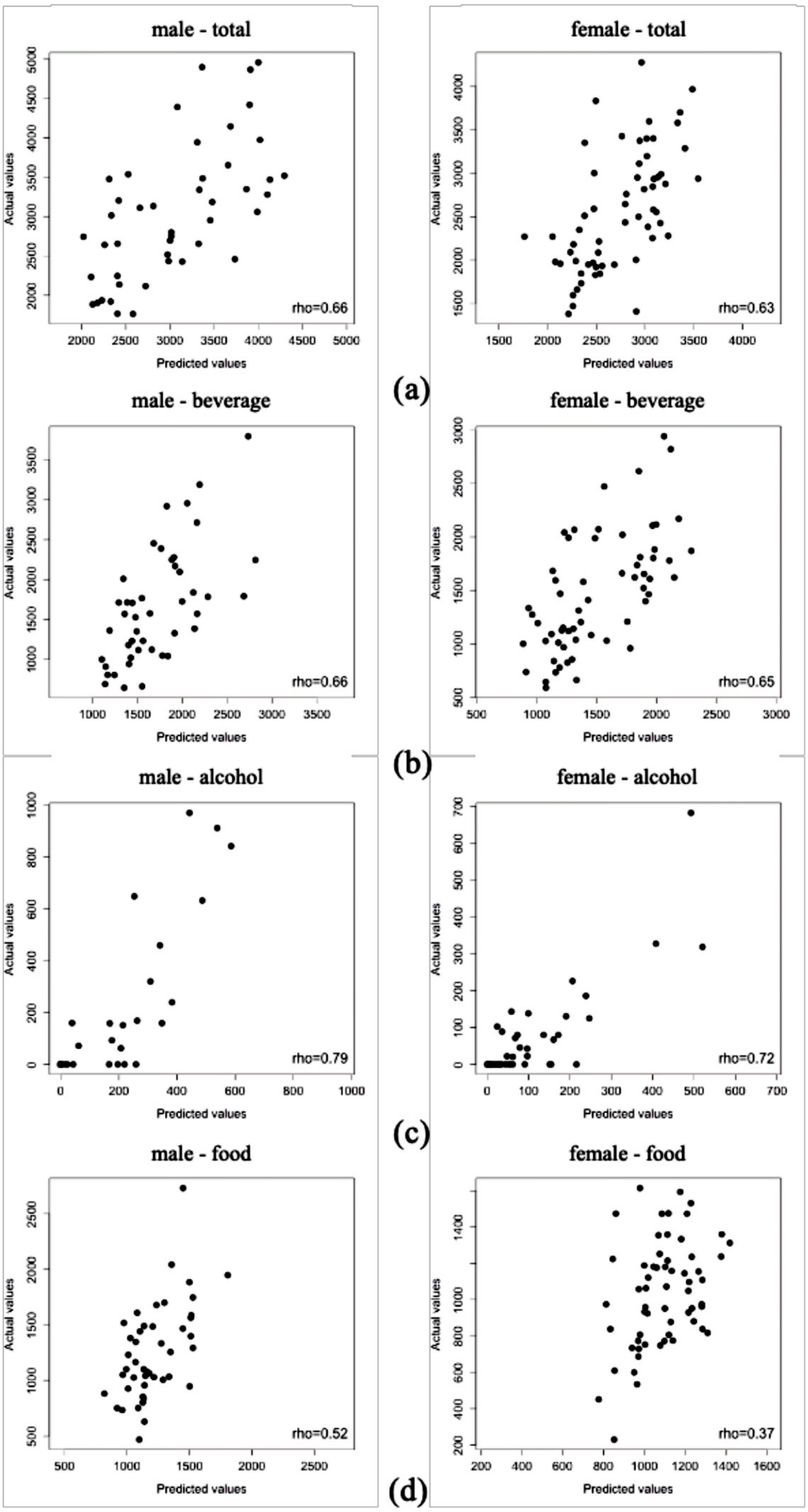


**Figure 3.** Prediction results were measured using the correlation coefficient of the best model of predicted and actual water intake values by the selective recall method in testing participants: water intake from total (**a**), beverage (**b**), alcohol (**c**), and food (**d**).

**Table III.** Spearman's correlation for total water intake after reducing ambiguity questions in crowdsourcing. The bold value represents the best correlation.

| | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 6 | Top 7 | Top 8 | Top 9 | Top 10 | Top 11 | Top 12 | Top 13 | Top 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Males | 0.14 | 0.47 | 0.51 | 0.43 | 0.50 | 0.46 | 0.48 | **0.62** | 0.52 | **0.62** | 0.61 | 0.54 | 0.53 | 0.53 |
| Females | 0.31 | 0.48 | 0.35 | 0.48 | 0.46 | 0.50 | 0.57 | 0.50 | 0.58 | 0.64 | 0.61 | **0.65** | 0.58 | 0.55 |

0.61 for top 9, 0.67 for top 9, and 0.33 for top 10 for beverage-derived, alcohol-derived, and food-derived respectively, for females. Thus, there were no noticeable differences between the conditions before and after excluding the questions. We proposed that the maximum number of questions for validation should be 15 for total water intake and 10 for other sources. Thus, after reducing the three questions, we finally analyzed a maximum of 9 to 14 questions as the SQW for total water intake and a maximum of 9 or 10 questions as the SQW for others (Table IV, **Supplementary Table I**).

### *Data Collection for Descriptive Dietary Records and Validation*

The target of the validation analysis was 97 males (age, mean: 58.6, SD=5.5) and 101 females (ages, M: 57.8, SD=5.4). We also obtained the datasets of winter (33 males and 33 females), spring and autumn (31 males and 33 females), and summer (33 males and 35 females).

Table V presents Pearson's correlations for each season among males and females. We confirmed correlation coefficients of more than 0.50, although this is slightly lower than the crowdsourcing results. For total water intake from SQW, the top 11 for males and the 14 for females showed the highest correlation. We confirmed that water intake from alcohol was relatively easier to predict, corroborating the crowdsourcing data. Since Spearman's correlation was used during the development of SQW, we also checked the correlation coefficient using Spearman's correlation coefficient, and there was almost no difference between Spearman's and Pearson's. Furthermore, correlations in the quintile value were confirmed for use in epidemiological analyses. Pearson's correlation coefficients for water intake to descriptive recall methods when the SQW values were in quantiles are 0.49, 0.54, 0.55, 0.49, 0.49, and 0.50, from top 9 to top 14 for males and 0.39, 0.42, 0.55, 0.51, 0.55, and 0.59 top 9 to top 14 for females. There was uncertainty about whether the study subjects represented the general Japanese population. However, although the obtained results contain some bias, they are likely to be somewhat generalizable because the data were collected on a sufficient sample size and with seasonality taken into account. SQW has been shown to provide a simple way to determine habitual water intake. It would help understand the distribution of water intake in epidemiology and elderly facilities, and so on.

## Discussion

Using the selective recall method, we predicted the habitual water intake based on a four-day average dietary recall, which revealed a good correlation between the predicted and water intake values. We further tested the models in the validation study and confirmed a Pearson's correlation of 0.5 in most cases. Our SQW took only 2-3 minutes for a total water intake or 8-9 minutes for each source and total.

The trained models in random forests can be understood in terms of their weighted features, which makes our question set interpretable. For example, drinking tea and water are questions that predict habitual water intake. These trained models and questionnaires can be used in validation studies with dietary records. Our detailed question lists for water intake from non-alcoholic beverages, alcohol, and food (other than total water intake) are available upon request. The food results showed a lower prediction range because the question set was limited. Food includes various intakes, including noodles, fruits, and vegetables. Therefore, having fewer questions complicates the estimation of water intake from food because it is necessary to cover a variety of foods. To predict water intake from food more precisely, we must broadly design new questions that specifically ask about food amount consumption. In this task, we confirmed that random forests

**Table IV.** Simplified lifestyle questionnaire for predicting habitual water intake (SQW) for total water intake. All questions were translated into English.

| Males | Question | Answer |
|---|---|---|
| 1 | How many times a day do you eat a meal? | l.t. once/once/twice/three times/m.t. four times |
| 2 | How much do you usually drink in the morning*1? | l.t. 200 mL/m.t. 200 mL/m.t. 350 mL/m.t. 500 mL/m.t. 700 mL/m.t. 1,000 mL/m.t. 1,500 mL |
| 3 | Of the staple foods, how many times a week do you eat sushi, takikomi-gohan, donburi, or okonomiyaki? | (Describe the number of times) |
| 4 | How much light-colored vegetables and mushrooms do you eat for breakfast? | Never/Small/Normal/Large |
| 5 | How many soybeans/soy products do you eat for breakfast at once? | Never/Small/Normal/Large |
| 6 | How many pickles do you eat at once? | Never/Small/Normal/Large |
| 7 | How much tea do you usually drink in a day? 1 cup: approximately 200 mL | Never/l.t.1 cup/1 cup/2 cups/3 cups/4-5 cups/6-7 cups/ m.t. 8 cups |
| 8 | How much do you usually drink in the afternoon?*2 | l.t. 200 mL/ m.t. 200 mL/m.t. 350 mL/m.t. 500 mL/m.t. 700 mL/m.t. 1,000 mL/m.t. 1,500 mL |
| 9 | How many eggs do you eat per week? | (Describe the number of eggs) |
| 10 | How many times a week do you eat salty foods? | (Describe the number of times) |
| 11 | How much soup do you drink when you eat noodles? | Almost all/80%/40-60%/20%/Almost none |
| 12 | How much green and yellow vegetables do you eat at a time for breakfast? | Never/Small/Normal/Large |
| 13 | How much soup is left at a time when you eat noodles? | Almost all/Approximately half/Almost none |
| 14 | How often do you usually eat noodles? | Everyday/5-6 days a week/3-4 days a week/1-2 days a week/1-3 day a month/Hardly |
| **Females** | **Question** | **Answer** |
| 1 | How much do you usually drink in the afternoon*2? | l.t. 200 mL/m.t. 200 mL/m.t. 350 mL/m.t. 500 mL/m.t. 700 mL/m.t. 1,000 mL/m.t. 1,500 mL |
| 2 | How much do you typically drink in the morning*1? | l.t. 200 mL/m.t.200 mL/m.t. 350 mL/m.t. 500 mL/m.t. 700 mL/m.t. 1,000 mL/m.t. 1,500 mL |
| 3 | How much carbonated beverages, fruit juices, and soft drinks do you usually drink daily? 1 cup: approximately 200 mL | Never/l.t.1 cup/1 cup/2 cups/3 cups/4-5 cups/6-7 cups/m.t. 8 cups |
| 4 | How much tea do you usually drink in a day? 1 cup: approximately 200 mL | Never/l.t.1 cup/1 cup/2 cups/3 cups/4-5 cups/6-7 cups/m.t. 8 cups |
| 5 | Do you think you drink at a faster pace compared to others? | Faster/Somewhat faster/Same/Somewhat slower/Slower |
| 6 | How much miso soup do you usually eat every day? 1 cup: about one bowl | Never/l.t.1 cup/1 cup/2 cups/3 cups/4-5 cups/6-7 cups/m.t. 8 cups |
| 7 | How many cups of miso soup do you drink in a week? | (Describe the number of cups) |
| 8 | How many times a week do you eat light-colored vegetables and mushrooms for breakfast? | (Describe the number of times) |
| 9 | Do you drink at mealtime? | Always/Almost always/Rarely/Never |
| 10 | Do you think you drink more than others? | More/Somewhat more/Same/Somewhat less/Less |
| 11 | How much light-colored vegetables and mushrooms do you eat at a time for lunch? | Never/Small/Normal/Large |
| 12 | Do you feel sick from cold food? | Never/Rarely/Sometimes/Often/Always |
| 13 | How much meat/processed meat do you eat for breakfast at one time? | Never/Small/Normal/Large |
| 14 | Choose one that is similar to your normal state regarding bowel movements. | Prone to diarrhea/somewhat diarrhea/Normal/ sometimes constipated/Constipated |

l.t.: less than, m.t.: more than. *1: morning is the time from waking up to before lunch. *2: afternoon is the time from lunch to dinner.

outperformed other regression algorithms in terms of interpretability and prediction. We also hypothesize that deep learning models can improve the accuracy of a model (high correlations).

### *Limitations*

The study includes a few limitations. The number of data samples was too small, despite augmenting the data. We believe that random forest and other algorithms are potentially helpful for

**Table V.** Pearson's correlations in seasons for validation: bold values represent the best correlation for total water intake.

| | | | N | Top 9 | Top 10 | Top 11 | Top 12 | Top 13 | Top 14 |
|---|---|---|---|---|---|---|---|---|---|
| Total | Average of a year | Males | 97 | 0.53 | 0.55 | **0.58** | 0.56 | 0.56 | 0.51 |
| | | Females | 101 | 0.40 | 0.45 | 0.53 | 0.53 | 0.54 | **0.58** |
| | Winter | Males | 33 | 0.46 | **0.55** | 0.52 | 0.51 | 0.47 | 0.43 |
| | | Females | 33 | 0.39 | 0.38 | 0.47 | 0.45 | 0.46 | **0.49** |
| | Spring or Autumn | Males | 31 | 0.62 | 0.65 | 0.65 | **0.69** | 0.72 | 0.66 |
| | | Females | 33 | 0.43 | 0.48 | 0.46 | 0.49 | 0.50 | **0.57** |
| | Summer | Males | 33 | 0.57 | 0.55 | **0.63** | 0.50 | 0.54 | 0.52 |
| | | Females | 35 | 0.47 | 0.53 | 0.67 | 0.67 | 0.66 | **0.71** |
| Beverage | Average of a year | Males | 97 | - | **0.39** | - | - | - | - |
| | | Females | 101 | **0.55** | - | - | - | - | - |
| Alcohol | Average of a year | Males | 97 | - | **0.61** | - | - | - | - |
| | | Females | 101 | **0.76** | - | - | - | - | - |
| Food | Average of a year | Males | 97 | **0.34** | - | - | - | - | - |
| | | Females | 101 | - | **0.34** | - | - | - | - |

- Indicates not available.

this prediction task. However, since some food frequency questionnaires used in epidemiology have correlation coefficients of approximately 0.3 with the gold standard[15], the SQW from food in this study can be usable in epidemiological analysis.

Compared to previous cohort studies[37], the results showing Spearman's rank correlation coefficients for the validity of the FFQ estimated water intake relative to water intake were 0.41 and 0.71. Furthermore, in other prior cohort studies[38], Spearman's correlation coefficients between non-alcoholic beverage consumption and dietary records obtained from the FFQ were 0.43 for men and 0.28 for women in one cohort participant and 0.56 and 0.58 for another cohort participant, respectively. Given the above, the correlation coefficients for the study in the SQW are comparable. The SQW is considered accurate enough to determine water intake in epidemiological studies. The SQW will make it possible to determine the approximate water intake of the Japanese population and deepen the analysis of the relationship between water intake and health status, thereby building scientific evidence related to fluid intake. In addition, by quickly ascertaining the water intake of each facility and population, it is thought that this information can be widely deployed in educational activities on water intake.

## Conclusions

This is the first study to create new questions, SQW, and evaluate the validity of examining the relationship between selective and descriptive recall methods and lifestyle-related questions. This study collected large-scale data to predict habitual water intake accurately, with at least nine or ten questions. We further validated the models in the validation study. We hypothesize that dissemination of SQW can lead to better health management by easily determining the habitual water intake.

### Conflict of Interest

The Authors declare that they have no conflict of interests.

### Funding

This study was founded by Suntory Holdings, Ltd. and Suntory Global Innovation Center, Ltd., Japan.

### Ethics Approval

This study was conducted following the guidelines in the Helsinki Declaration (revised by the Fortaleza General Meeting of the World Medical Association). All procedures involving human participants were approved by the ethics committee of Suntory Holdings Limited (IRB No. 15000139) and Nara Institute of Science and Technology (IRB No. 2019-I-19). This study also completed with the Ethical Guidelines for Medical Research Involving Human Subjects (2014 Ministry of Education, Culture, Sports, Science and Technology and the Ministry of Health, Labour and Welfare Ministerial Notification No. 3). All procedures involving human participants were approved by the Ethics Committee of the Foundation for Biomedical Research and Innovation at Kobe (IRB No. 20000109) and the ethics committee of Miura Clinic, Medical Corporation Kanonkai (IRB No. 17000161).

### Informed Consent

Informed consent was obtained from all individual participants included in the study.

### Authors' Contribution

Conceptualization: Ken Suzuki, Harumichi Seta, Hiroshi Watanabe, Toshihide Suzuki, Reiji Watanabe, Norihito Murayama, Tomonori Okamura, Satoshi Nakamura. Data curation: Hiroki Tanaka, Aya Hirata, Junji Miyazaki. Investigation: Hiroki Tanaka, Aya Hirata, Junji Miyazaki, Keiko Mizuma, Yumi Nakamura. Project administration: Ken Suzuki, Harumichi Seta, Hiroshi Watanabe, Toshihide Suzuki, Reiji Watanabe, Norihito Murayama, Tomonori Okamura, Satoshi Nakamspearura. Writing original draft: Hiroki Tanaka. Writing-review and editing: Keiko Mizuma, Yumi Nakamura, Aya Hirata, Junji Miyazaki, Ken Suzuki, Harumichi Seta, Hiroshi Watanabe, Toshihide Suzuki, Reiji Watanabe, Norihito Murayama, Tomonori Okamura, Satoshi Nakamura. Supervisor: Tomonori Okamura, Satoshi Nakamura.

### Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

### ORCID ID

Hiroki Tanaka: 0000-0002-0548-6252
Yumi Nakamura: 0000-0002-5915-0498
Aya Hirata: 0000-0003-1439-4242
Junji Miyazaki: 0000-0003-3634-6401
Toshihide Suzuki: 0000-0003-4796-0749
Norihito Murayama: 0000-0003-2895-7106
Tomonori Okamura: 0000-0003-0488-0351
Satoshi Nakamura: 0000-0001-6956-3803

## References

1) Kleiner SM. Water: an essential but overlooked nutrient. J Am Diet Assoc 1999; 99: 200-206.
2) J'equier E, Constant F. Water as an essential nutrient: the physiological basis of hydration. Eur J Clin Nutr 2010; 64: 115-123.
3) Haussinger D, Lang F, Gerok W. Regulation of cell function by the cellular hydration state. Am J Physiol Endocrinol Metab 1994; 267: 343-355.
4) Cotter JD, Thornton SN, Lee JK, Laursen PB. Are we being drowned in hydration advice? Thirsty for more? Extrem Physiol Med 2014; 3: 1-16.
5) Nishikawa T, Miyamatsu N, Higashiyama A, Hojo M, Nishida Y, Fukuda S, Hirata T, Ichiura K, Kubota Y, Kubo S, Ueba T, Kadota A, Sugiyama D, Okamura T. Daily habit of water intake in patients with cerebral infraction before its onset; Comparison with a healthy population: a cross-sectional study. Cerebrovasc Dis 2019; 47: 143-150.
6) Armstrong LE, Ganio MS, Casa DJ, Lee EC, McDermott BP, Klau JF, Jimenez L, Le Bellego L, Chevillotte E, Lieberman HR. Mild dehydration affects mood in healthy young women. J Nutr 2012; 142: 382-388.
7) Krecar I, Kolega M, Kunac S. The Effects of Drinking Water on Attention. Procedia Soc Behav Sci 2014; 159: 577-583.
8) Secher M, Ritz P. Hydration and cognitive performance. J Nutr Health Aging 2012; 16: 325-329.
9) Ritz P, Berrut G. The importance of good hydration for day-to-day health. Nutr Rev 2005; 63: 6-13.
10) Kleiner SM. Water: an essential but overlooked nutrient. J Am Diet Assoc 1999; 99: 200-206. Erratum in: J Am Diet Assoc 1999; 99: 411.
11) Tanaka H, Iwasaka H, Negoro H, Nakamura S. Analysis of conversational listening skills toward agent-based social skills training. J Multimodal User Interfaces 2020; 14: 73-82.
12) Tanaka H, Adachi H, Ukita N, Ikeda M, Kazui H, Kudo T, Nakamura S. Detecting Dementia Through Interactive Computer Avatars. IEEE J Transl Eng Health Med 2017; 5: 1-11.
13) Maruyama K, Kokubo Y, Yamanaka T, Watanabe M, Iso H, Okamura T, Miyamoto Y. The reasonable reliability of a self-administered food frequency questionnaire for an urban, Japanese, middle-aged population: the Suita study. Nutr Res 2015; 35: 14-22.
14) Date C, Fukui M, Yamamoto A, Wakai K, Ozeki A, Motohashi Y, Y, Adachi C, Okamoto N, Kurosawa M, Tokudome Y, Kurisu Y, Watanabe Y, Ozasa K, Nakagawa S, Tokui N, Yoshimura T, Tamakoshi A and JACC Study Group. Reproducibility and Validity of a Self-administered Food Frequency Questionnaire Used in the JACC Study. J Epidemiol 2005; 15: 9-23.
15) Okada C, Iso H, Ishihara J, Maruyama K, Sawada N, Tsugane S. Validity and reliability of a self-administered food frequency questionnaire for the JPHC study: The assessment of amino acid intake. J Epidemiol 2017; 27: 242-247.
16) Ghezzi A, Gabelloni D, Martini A, Natalicchio A. Crowdsourcing: A Review and Suggestions for Future Research. Int J Manag Rev 2018; 20: 343-363.
17) Tokuda S, Tanaka H, Nakamura S, Murayama N, Kyo H, Uebara K, Suzuki N, Kanaya S. A Cross-sectional Study on the Association between the Constitution in Chinese Medicine and the risk factors of Lifestyle Diseases. Jpn J Complement Altern Med 2019; 16: 95-104.
18) Yamamori A, Kyo H, Watanabe T, Huang M, Ono N, Sato T, Abe T, Uebaba K, Kawabata K, Imanishi K, Altaf-Ul-Amin M, Zhu Y, Dai Z, Wang Q, Kanaya S, Ohta T, Suzuki N. Relationship between 60 Items in Japanese Version of the Constitution in Chinese Medicine Questionnaire (CCMQ-J) Based on Multivariate Analysis: Estimation of Aging and BMI by CCMQ-J Scores. Jpn J Complement Altern Med 2016; 13: 43-56.

19) Altunkaynak A, Assefa T. Monthly Water Consumption Prediction Using Season Algorithm and Wavelet Transform-Based Models. J Water Resour Plan Manag 2017; 143.

20) Sasaki S, Takahashi T, Iitoi Y, Iwase Y, Kobayashi M, Ishihara J, Tsugane S and JPHC Study Group. Food and Nutrient Intakes Assessed with Dietary Records for the Validation Study of a Self-administered Food Frequency Questionnaire in JPHC Study Cohort I. J Epidemiol 2003; 13: 23-50.

21) Murakami M, Mizuma K, Nakamura Y, Watanabe R. Estimation of water intake from food moisture in the Japanese diet using cooking-based conversion factor for water content. J Food Sci 2021; 86: 266-275.

22) Yoshimura Y, Takahashi K. Excel Eiyho-kun Food Frequency Questionnaire based on food group FFQg (computer manual and software). Tokyo: Kenpakusha (in Japanese), 2014.

23) Nakatsu N, Sawa R, Misu S, Ueda Y, Ono R. Reliability and validity of the Japanese version of the simplified nutritional appetite questionnaire in community-dwelling older adults. Geriatr Gerontol Int 2015; 15: 1264-1269.

24) Doi Y, Minowa M, Uchiyama M, Okawa M, Kim K, Shibui K, Kamei Y. Psychometric assessment of subjective sleep quality using the Japanese version of the Pittsburgh Sleep Quality Index (PSQI-J) in psychiatric disordered and control subjects. Psychiatry Res 2000; 97: 165-172.

25) Comprehensive Survey of Living Conditions. Available from: https://www.mhlw.go.jp/english/database/db-hss/cslc.html.

26) Zhu Y, Origasa H, Uebara K, Xu F, Wang Q. Development and Validation of the Japanese Version of the Constitution in Chinese Medicine Questionnaire (CCMQ). Kampo Med 2008; 59: 783-792.

27) Iwasa H, Yoshida Y. Psychometric evaluation of the Japanese version of Ten-Item Personality Inventory (TIPI-J) among the middle-aged and elderly adults: Concurrent validity, internal consistency and test-retest reliability. Cogent Psychol 2018; 5: 465-476.

28) Oshio A, Abe S, Cutrone P. Development, Reliability, and Validity of the Japanese Version of Ten Item Personality Inventory (TIPI-J). Jpn J Personal 2012; 21: 40-52.

29) Uchida T, Nakamura Y, Tanaka H, Nakamura S, Okamura T, Watanabe H, Murayama N. Validity of a selective recall method for assessing water intake and its relationship with hydration status. Eur Rev Med Pharmacol Sci 2021; 25: 6623-6632.

30) Tani Y, Asakura K, Sasaki S, Hirota N, Notsu A, Todoroki H, Miura A, Fukui M, Date C. The influence of season and air temperature on water intake by food groups in a sample of free-living Japanese adults. Eur J Clin Nutr 2015; 69: 907-913.

31) Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, Del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. Array programming with NumPy. Nature 2020; 585: 357-362.

32) R Core Team. R: A Language and Environment for Statistical Computing. Available from: http://www.R-project.org/.

33) Torgo L, Branco P, Ribeiro RP, Pfahringer B. Resampling Strategies for Regression. Expert Sys 2015; 32: 465-476.

34) Breiman L. Random Forests. Machine Learning 2001; 45: 5-32.

35) Umemoto K, Kubo S, Nishida Y, Higashiyama A, Kawamura K, Kubota Y, Hirata T, Hirata A, Sata M, Kuwabara K, Miyazaki J, Kadota A, Iida M, Sugiyama D, Miyamatsu N, Miyamoto Y, Okamura T. Physique at birth and cardiovascular disease risk factor in Japanese urban residents: the KOBE study. J Atheroscler Thromb 2022; 29: 188-199.

36) Kato E, Takachi R, Ishii Y, Sasazuki S, Sawada N, Iwasaki M, Shinozawa Y, Umezawa J, Tanaka J, Yokoyama Y, Kitamura K, Nakamura K, Tsugane S. Online version of the self-administered food frequency questionnaire for the Japan Public Health Center-based Prospective Study for the Next Generation (JPHC-NEXT) protocol: Relative validity, usability, and comparison with a printed questionnaire. J Epidemiol 2017; 27: 435-446.

37) Cui R, Iso H, Echak ES, Murayama K, Tamakoshi A and JACC study Group. Water intake from foods and beverages and risk of mortality from CVD: the Japan collaborative cohort (JACC) study. Public Health Nutr 2018; 21: 3011-3017.

38) Cui R, Iso H, Eshak ES, Maruyama K, Ishihara J, Takachi R, Sawada N, Tsugane S and JPHC Study Group. Non-alcoholic beverages intake and risk of CVD among Japanese men and women: the Japan public health center study. Br J Nutr 2022; 127: 1742-1749.