Predicting habitual water intake from lifestyle questions

H. TANAKA¹, K. MIZUMA², Y. NAKAMURA³, A. HIRATA⁴, J. MIYAZAKI⁴, K. SUZUKI², H. SETA², H. WATANABE³, T. SUZUKI³, R. WATANABE², N. MURAYAMA³, T. OKAMURA⁴, S. NAKAMURA¹

¹Division of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan
²Suntory Global Innovation Center Limited, Institute for Water Science, Seika-cho, Soraku-gun, Kyoto, Japan
³Suntory Global Innovation Center Limited, Research Institute, Seika-cho, Soraku-gun, Kyoto, Japan
⁴Department 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.

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 body¹⁻². 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 substances³. 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 disorders⁴, and is associated with the onset of cerebral infarction⁵. 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 headaches⁶⁻⁷. For example, Secher and Ritz⁸ established a clear link between dehydration and reduced cognitive performance. Other studies⁹⁻¹⁰ showed that drinking more water may improve performance
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\textsuperscript{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\textsuperscript{13,14} and specific amino acid intake\textsuperscript{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\textsuperscript{6,17}. Partial least squares (PLS) regression and factor analysis were used in a previous study\textsuperscript{18} to predict age and body mass index and to observe question weights. Another previous study\textsuperscript{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\textsuperscript{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\textsuperscript{48} 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\textsuperscript{22} since the frequency of food intake may be affected. Other questions were from the simplified nutritional appetite questionnaire (SNAQ-J)\textsuperscript{23} and Pittsburgh Sleep Assessment\textsuperscript{24}, which calibrates sleep quality and smoking behaviors. We also obtained questionnaires from the National Health and Nutrition Survey\textsuperscript{25} and the Japanese version of the Constitution in the Chinese Medicine Questionnaire (CCMQ)\textsuperscript{26}. We also adopted the Ten Item Personality Inventory (TIPI-J)\textsuperscript{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
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 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.

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. 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 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, 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; 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: \( d_{post} \) \{2,3,4,5,10,15\} for post-feature selection, and \( d_{pre} \) \{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.](image-url)
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 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 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. 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 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 (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.

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 from 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.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Beverage</th>
<th>Alcohol</th>
<th>Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>2,998.5 (1,017.7)</td>
<td>1,658.3 (773.3)</td>
<td>123.2 (221.7)</td>
<td>1,217.0 (479.1)</td>
</tr>
<tr>
<td>Females</td>
<td>2,566.1 (735.7)</td>
<td>1,439.8 (578.5)</td>
<td>56.5 (141.1)</td>
<td>1,069.7 (319.2)</td>
</tr>
</tbody>
</table>
Predicting habitual water intake from lifestyle questions

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.

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

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

<table>
<thead>
<tr>
<th></th>
<th>Top 1</th>
<th>Top 2</th>
<th>Top 3</th>
<th>Top 4</th>
<th>Top 5</th>
<th>Top 6</th>
<th>Top 7</th>
<th>Top 8</th>
<th>Top 9</th>
<th>Top 10</th>
<th>Top 11</th>
<th>Top 12</th>
<th>Top 13</th>
<th>Top 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>0.14</td>
<td>0.47</td>
<td>0.51</td>
<td>0.43</td>
<td>0.50</td>
<td>0.46</td>
<td>0.48</td>
<td>0.62</td>
<td>0.52</td>
<td>0.62</td>
<td>0.61</td>
<td>0.54</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>Females</td>
<td>0.31</td>
<td>0.48</td>
<td>0.35</td>
<td>0.48</td>
<td>0.46</td>
<td>0.50</td>
<td>0.57</td>
<td>0.50</td>
<td>0.58</td>
<td>0.64</td>
<td>0.61</td>
<td>0.65</td>
<td>0.58</td>
<td>0.55</td>
</tr>
</tbody>
</table>

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

Compared to previous cohort studies\cite{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\cite{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. 2019-1-19). This study also complied 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, Satoshi Nakamura, Tomonori Okamura. Writing-original draft: Hiroki Tanaka.

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: 0000-0003-6364-6401
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
Predicting habitual water intake from lifestyle questions


