frontiers

Soft tap water urgently needed for reducing risks of kidney stones at rural villages in Yangxin, a poverty alleviated county in central China

Jiaxin Zhao[1], Mingyao Wang[1], Tan Jiang[1], Fangsi Wang[1], Xinyue Shi[2], Yun Zhang[1], Kun Xu[1*]

[1] College of Life Sciences, Hubei Normal University, 11 Cihu Road, Huangshi, Hubei, China

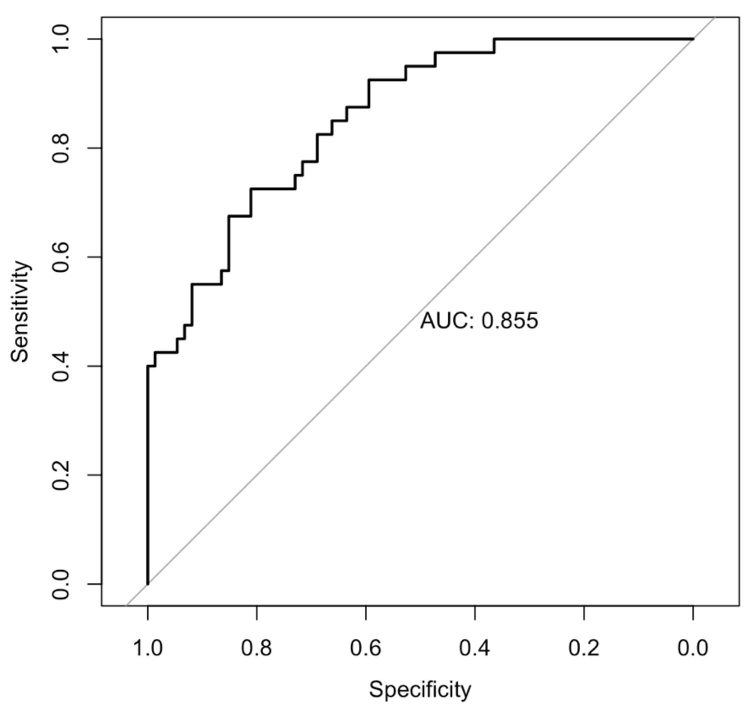
[2] Center for Public Employment and Entrepreneurship Guidance and Information Service, Department of Human Resources and Social Security of Hubei Province, 10 Shuiguohu Road, Wuhan, Hubei, China

*** Correspondence:**
Kun Xu
kunxu1@hbnu.edu.cn

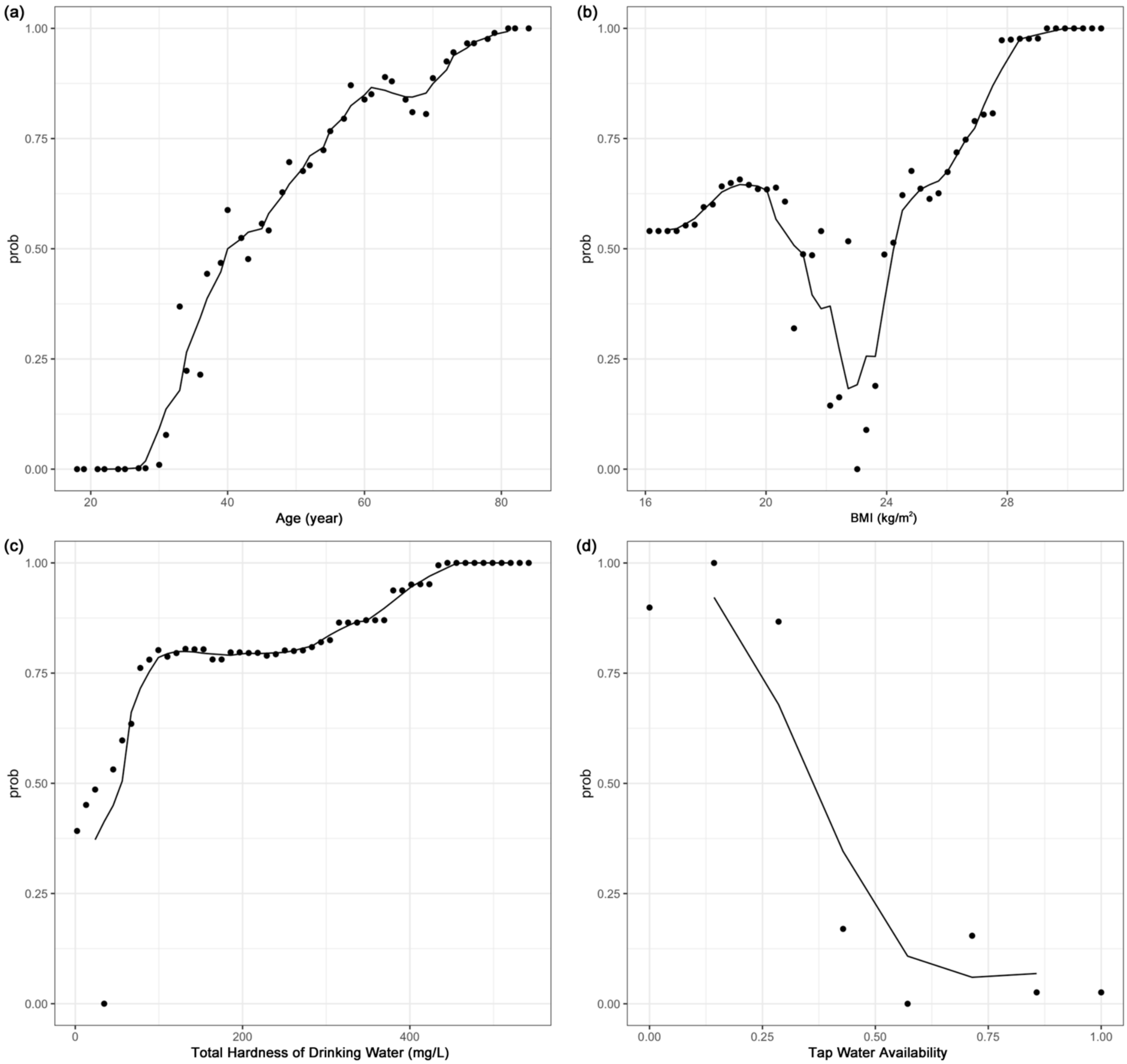
Supplementary Material

1 Supplementary Figures and Tables

1.1 Supplementary Figures



Supplementary Figure 1. Receiver Operating Characteristic (ROC) curve of the estimated logistic regression model with an area under curve (AUC) of 0.855.



Supplementary Figure 2. Partial dependence of prevalence of kidney stones (prob) on (a) age (year), (b) body mass index (BMI) (kg/m^2), (c) total hardness of drinking water (mg/L $CaCO_3$), and (d) tap water availability based on the random forest regression marginalizing over other variables. Solid line was the mean estimated probability at the rolling window of 3 adjacent estimates (black dots).

1.2 Supplementary Tables

Supplementary Table 3. Questions answered by householders during family visits regarding demographic (1-6), economic (7), and urinary health (8) conditions, water availability (9-15), and perspectives on water safety (16-17) and government support (18-22).

No	Question	Answer Format (Unit)
1	Your name?	Text
2	Your gender?	Female/Male
3	Your age?	Integer (year)
4	Your height?	Numeric (m)
5	Your weight?	Numeric (kg)
6	Number of adults in your family?	Integer
7	Annual household income?	Numeric (CNY)
8	Are you diagnosed with kidney stones at hospital?	No/Yes
9	Is there a well in your family yard?	No/Yes
10	Is there a filter installed to the well pump?	No/Yes (Blank if No to Question 9)
11	Is tap water available in your family?	No/Yes
12	How many days in the last 3 months was tap water available for?	Numeric (day) (0 if No to Question 11)
13	Is there a filter installed at the kitchen faucet?	No/Yes
14	How much did your family pay for water last year?	Numeric (CNY)
15	Your primary source of drinking water?	Well/Filtered Well/Tap/Filtered Tap/Bottle Water
16	Are you aware of any water issues in the recent 3 years?	No/Yes
17	Have you noticed deterioration of water quality during the past 3 years?	No/Yes
18	Are you willing to accept compensation from government regarding water problems?	No/Yes
19	How much are your family willing to accept each year?	Numeric (CNY) (Blank if No to Question 18)
20	Are you confident in government regarding compensation for water problems?	No/Yes
21	Do you hope government to treat water problems?	No/Yes
22	Are you willing to pay government for water treatment?	No/Yes

Supplementary Table 2. Performance of the logistic regression model built on a reduced size of survey data. Simulation for dropping 10% of responses at each village was conducted for 1000 times. Estimated coefficients and p value of each coefficient were summarized.

Variable	Mean Estimated coefficient	Difference from the fitted model (Table 4)	Number of times with p value < 0.05
(Intercept)	-8.21	0.04	1000
Gender	1.34	-0.01	923
Age	8.92×10^{-2}	-0.09×10^{-2}	1000
BMI^2	4.47×10^{-3}	0.03×10^{-3}	574
Total hardness of drinking water	9.64×10^{-3}	0.07×10^{-3}	997
Tap water availability	-2.23	0.02	719

1.3 Supplementary R Codes

```
library(MASS)
library(pROC)
library(randomForest)
library(ggplot2)
library(pdp)

Water<-read.csv("./Water.csv") #. is the directory path

#Building a full logistic regression model
binomial1<-
 glm(Self_Stones~Drinking_tws+CaMg_Ratio+Softening_Depth+Tap_Availab
 ility+Fee_Clean_YR+Gender_M+BMI+BMI2+Family_Income+percapita_Income
 +Aware_Water_Issue+Willingness_To_Accept+Willingness_To_Pay+Hope_Go
 v_Treat+Confidence_Gov, data=Water, family=binomial(link="logit"))

#stepwise selection based on AIC
binomial2<-stepAIC(binomial1)

summary(binomial2)
#This leads to Table 4

rr<-roc(Water$Self_Stones, predict(binomial2))

#AUC of the ROC of the logistic regression model
rr$auc

#plotting the ROC curve
plot(roc(Water$Self_Stones, predict(binomial2)))


#Uncertainty Analysis by randomly dropping 10% of residents at each
```

```
 village by 1000 times of iteration

coef_1000<-matrix(nrow=6, ncol=1000)

p_1000<-matrix(nrow=6, ncol=1000)

for(i in 1:1000)
{
set.seed(i)
Water_Drop_10<-c(sample(21,2), sample(42,4)+21, sample(24,2)+63,
 sample(27,3)+87)

Water_Drop_10_Data<-Water[-Water_Drop_10,]
glm_Drop_10<-
 glm(Self_Stones~Gender_M+Age_YR+I(BMI^2)+Drinking_tws+Tap_Availabil
 ity, Water_Drop_10_Data, family=binomial(link="logit"))

coef_1000[,i]<-summary(glm_Drop_10)$coefficients[,1]

p_1000[,i]<-summary(glm_Drop_10)$coefficients[,4]
}

#mean estimated coefficients from 1000 times of iteration of random
 dropping 10% samples
apply(coef_1000,1,mean)

#significance of each variable by 1000 times of iteration
apply(p_1000<0.05,1,sum)


#Random forest regression model for the binary response
Water$Stones<-as.factor(Water$Self_Stones)

set.seed(691)
RF<-randomForest(Stones ~ Gender_M + Age_YR + BMI + Drinking_tws +
 Tap_Availability, data = Water, ntree = 300)

#variance importance plot
varImpPlot(RF)

#partial dependence on gender
partial(RF, pred.var="Gender_M")

#plotting partial dependence on each continuous explanatory
 variable
autoplot(partial(RF, pred.var="Age_YR"))+theme_bw()

autoplot(partial(RF, pred.var="BMI"))+theme_bw()
```

```
autoplot(partial(RF, pred.var="Drinking_tws"))+theme_bw()

autoplot(partial(RF, pred.var="Tap_Availability"))+theme_bw()

#plotting partial dependence on the combination of two continuous
 variables
autoplot(partial(RF, pred.var=c("Age_YR","BMI")))+theme_bw()

autoplot(partial(RF,
 pred.var=c("Drinking_tws","Tap_Availability")))+theme_bw()
```