Original Article

Age at Natural Menopause; A Data Mining Approach (Data from the National Health and Nutrition Examination Survey 2013-2014)

Tahereh Alinia, Soheila Khodakarim[1], Fahimeh Ramezani Tehrani[2], Siamak Sabour[3]

Student Research Committee, School of Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [1]Department of Epidemiology, School of Allied Medical Sciences, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [2]Reproductive Endocrinology Research Center, Research Institute for Endocrine Sciences, Tehran, Iran, [3]Department of Clinical Epidemiology, School of Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, I.R. Iran

Abstract

Background: The timing of the age at which menopause occurs varies among female populations. This variation is attributed to genetic and environmental factors. This study aims to investigate the determinants of early and late-onset menopause. **Methods:** We used data from the National Health and Nutrition Examination Survey 2013-2014 for 762 naturally menopause women. Data on sociodemographic, lifestyle, examination, and laboratory characteristics were examined. We used random forest (RF), support vector machine (SVM), and logistic regression (LR) to identify important determinants of early and late-onset menopause. We compared the performance of models using sensitivity, specificity, Brier score, and area under the receiver operating characteristic (AUROC). The top determinants were assessed by using the best performing models, using the mean decease in Gini. **Results:** Random forest outperformed LR and SVM with overall AUROC 99% for identifying related factors of early and late-onset menopause (Brier score: 0.051 for early and 0.005 for late-onset menopause). Vitamin B12 and age at menarche were strongly related to early menopause. Also, methylmalonic acid (MMA), vitamin D, body mass index (BMI) were among the top highly ranked factors contributing to early menopause. Features such as age at menarche, MMA, sex hormone-binding globulin (SHBG), BMI, vitamin B12 were the most important covariate for late-onset menopause. **Conclusions:** Menarche age and BMI are among the important contributors of early and late-onset menopause. More research on the association between vitamin D, vitamin B12, SHBG, and menopause timing is required which will produce invaluable information for better prediction of menopause timing.

Keywords: *Data mining, menopause, nutrition surveys*

Introduction

Concurrent to chronological aging both the number and quality of the oocytes in the ovaries decrease, the ovaries stop producing estrogen and progesterone, and consequently, the menstrual periods stop permanently.[1] Twelve consecutive months of menstruation cessation, for which there is no other obvious pathological or physiological cause than the loss of ovarian follicular activity is defined as natural menopause.[2]

The mechanism underlying the age at natural menopause (ANM) has not been completely understood.[3] Different genetic, social, and environmental factors are likely to be associated with variability in ANM, huge controversies exist and no established risk factor is documented.[4]

Considerable long-term adverse health implication has been reported for early menopause. Early menopause links independently with increased odds of rheumatoid arthritis.[5] The risk of death among women with early menopause is higher.[6,7] Late-onset menopause also carries health risks. It is a proxy of prolonged exposure to estrogen and a large number of ovulation, which consequently puts women at a higher risk of ovarian, breast, and endometrial cancer.[8]

Data mining utilizes statistical methods for data classification. These techniques have been frequently applied to epidemiologic data to classify determinants of health and have indicated a higher accuracy than classical methods.[9-11] Support vector machine (SVM), random forest (RF), logistic regression (LR) have been broadly utilized in this era. They are the most frequently used supervised learning methods for analyzing complex survey data.

Several studies have looked to identify significant risk factors of early and

How to cite this article: Alinia T, Khodakarim S, Ramezani Tehrani F, Sabour S. Age at natural menopause; a data mining approach (Data from the National Health and Nutrition Examination Survey 2013-2014). Int J Prev Med 2021;12:180.

Address for correspondence:
Dr. Siamak Sabour, Department of Clinical Epidemiology, School of Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, I.R. Iran.
E-mail: sabour@sbmu.ac.ir

Access this article online

Website: www.ijpvmjournal.net/www.ijpm.ir

DOI: 10.4103/ijpvm.IJPVM_647_20

Quick Response Code:

late-onset menopause. To the best of our knowledge, there is a lack of studies for ANM determinants using data mining methods. We hypothesized that the use of flexible, optimized data mining approaches on a large data set with many features would yield accurate classification and generate new invaluable hypotheses. In this study, we aim to identify significant determinants of early and late-onset menopause using data mining algorithms, and data from the national health and nutrition examination survey (NHANES) 2013-2014.

Methods

Data source

Using the NHANES 2013-2014, the health information on naturally menopause women was collected. NHANES is a cross-sectional survey conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of adults and children in the United States of America. It has a complex, four-stage sampling scheme that combines interviews, physical examinations, and laboratory tests of approximately 5000 non-institutionalized civilian resident population of the United States annually.

The menopause status was determined by asking women the question "Have you had regular periods in the past 12 months?" and if the answer was no, the next question was "What is the reason that you have not had regular periods in the past 12 months?". Answer choices were pregnancy, breastfeeding, hysterectomy, menopause, and other. Women were referred as naturally menopause if the cause of the lack of menstruation was stated as menopause. 762 women were included in this study who were naturally menopause and reported their ANM. Women were categorized into three strata by their ANM (early, timely, and late-onset menopause). Early menopause was defined as amenorrhea before the age of 45.[12] Timely menopause was defined as women having menopause between ages 45 to 55, and late-onset menopause was defined as if it has not begun until 55.[13]

Input features

Several hundred variables are available in the NHANES data sets. Variables were selected based on existing research looking at the determinants of ANM. Even variables with a possible relation with ANM were considered. Variables were dropped from the dataset if it was only available on subsamples instead of the whole NHANES sample or it had a high level of missing value (more than 35%). A total of 38 variables were included in the models. In our application, considered variables cover socio-demographic (e.g. Education level, race, marital status, ratio of family income to poverty), lifestyle (e.g. drinking), reproductive (e.g. history of prior pregnancy), examination (e.g. anthropometrics), and laboratory (e.g. vitamin D level) characteristics [Tables 1 and 2] Family income-to-poverty ratio is an index of socioeconomic standing and represents family income by poverty level.

Data imputation

Tackling the missing values is a prerequisite for applying data mining algorithms. The proportion of missing data ranged from 0% to 15% for all features but, "the number of alcohol drinks over the past 12 months", "history of Cocaine/heroin/methamphetamine use", "age at first live birth", "history of vaginal, anal, or oral sex", and "age at first sex" which 25 to 35% of data were missing. Missing values in the covariates were imputed with multiple imputations using the package MICE in R with 5 imputations and 50 iterations.

Class unbalance

Data sets in this study were class-imbalanced. Since the data set shows a large number of timely menopausal women, the unbalanced distribution of the variable classes influences the model's performance. An approach to combat this challenge is oversampling. With oversampling, we duplicated samples from the minority class. We rebalanced the data by oversampling the minority class (early or late-onset menopause) and then proceed with learning the classification model on balanced data.

Once the data have been imputed and balanced, data mining algorithms were run on each of the balanced imputed dataset, and then the estimates from each dataset were combined to obtain the final result.

Data mining methods

We used three supervised learning, including random forest (RF), support vector machines (SVM), and logistic regression (LR) for the classification of naturally menopause women to early or late-onset menopause. An independent model was created for early vs. timely menopause and late-onset vs. timely menopause.

The RF is a supervised ensemble classification model that grows many classification trees built from a random subset of features and bootstrap samples. The RF ensemble the prediction from each tree through voting. The most important parameters for RF after parameter tuning in our study were ntree = 100 denotes the number of trees in the forest. The parameter mtry = 6 (square root of the total number of variables) denotes the number of features randomly selected as candidate features at each split.

SVM is a supervised data mining model that uses classification algorithms for two-group classification problems. The SVM is based on mapping data to a higher dimensional space through a linear kernel function and choosing the maximum-margin hyper-plane that separates data. Thus, the goal of the SVM is to improve accuracy by the optimization of space separation. The SVM model trains the characteristics associated with ANM groups. The regularization parameter was considered 10.

LR is a supervised data mining model that is used to model a binary dependent variable. It uses the logit function to predict certain class probabilities. The prediction from a logistic regression model can be interpreted as the probability that the label is 1. It's always best to predict class probabilities instead of predicting classes.

Performance evaluation

The dataset was divided into a training set (80% of total samples) used to develop the classification models and a validation set (20% of total samples) used to assess the classification accuracy of each model. Model on the testing dataset was evaluated using performance statistics in terms of sensitivity, specificity, Brier score, and Area Under Receiver Operating Characteristic (AUROC). A Brier score is a way to verify the accuracy of a classification model. It is calculated as the mean squared differences of actual results and forecast probability. It ranges between 0 and 1. The lower Brier score indicates higher accuracy. AUROC is known as a global measure of classifier performance that provides a comprehensive measure to summarize the false-positive rate, or 1–specificity versus sensitivity of the classification method. AUROC demonstrates how well the early, and late-onset menopause can be classified by the algorithm.

The best data mining model was selected based on performance metrics. Within the best performing model, variable importance measures rank the variables concerning their relevance for classification. It is assessed by the Gini impurity criterion index. Mean Decrease in Gini is the average of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each decision tree in the random forest. A low Gini (i.e., a greater decrease in mean Gini) means that a particular variable contributes a greater role in classification and have more relevance to menopause timing. The criterion for selecting the most important variable was the five variables with the largest decrease in the mean Gini among each model. All analyses were run using R release 4.0.3.

Results

Table 1 presents the socio-demographic characteristic of the study population stratified by their ANM category. Out of 762 naturally menopause women, early menopause reported among 132 (17.3%), timely menopause among 575 (75.4%), and late-onset menopause among 55 (7.2%). The average age of women at study time was 62.8 ± 10.6 for early menopause, 64.1 ± 9.1 for timely, and 67.4 ± 6.3 for late-onset menopause (p-value = 0.009). There was a decline across the years since the final menstrual period (FMP) from early to late-onset menopause. The average years since FMP at the time of study among early, timely, and late-onset menopause women were 23.1 ± 11.5, 14.1 ± 9.2, and 9.4 ± 5.9, respectively. The ratio of family income to poverty was the lowest among early menopause women (p-value = 0.019). Description of input variables into data mining models were presented in Table 2.

Table 3 describes the comparative performance of three data mining models (RF, SVM, and LR) for the classification of menopause timing. The full set of 38 variables were used for training the classification models (number of features = 38). Five classification performance estimates were produced (one estimate per each imputed dataset), separately for early and late-onset menopause, and then one combined estimate was provided for each ANM group.

Within the early vs. timely menopause women, we found that RFs outperformed LRs and SVMs. AUROC ranged from 98.1 to 98.4% among imputed datasets in the RF model; combined AUROC was 99%. The Brier score for combined RF was the lowest value among models (0.051 for RF versus 0.200 for LR and 0.211 for SVM). The worst performing model was SVM with a combined AUROC at 68.6%.

In the classification of late-onset vs. timely menopause, the RF was the best discriminating classifier with an AUROC score of 99% among all imputed, and combined estimates, followed by LR and SVM each reported combined AUROC of 84.1% and 78.9, respectively. The lowest Brier score belonged to RF balanced (Brier score at 0.005 for combined data) and the highest score to SVM balanced (Brier score at 0.200 for combined data).

Figures 1 and 2 highlight the relative importance of variables by the RF model. As RF was the top-performing model, the mean decrease in Gini was used to compare the importance between the variables within the model. The criterion for selecting the most important variable was the five variables with the largest decrease in the mean Gini among each model.

Analysis of features in early menopause women showed vitamin B12 and age at menarche were the most important features which contribute substantially towards the classification of the RF model. Features including Methylmalonic acid (MMA), vitamin D (25OHD2 + 25OHD3), body mass index (BMI) were among top highly ranked variables contributing to the classification into early menopause.

The late-onset menopause data analysis suggests features such as age at menarche, MMA, sex hormone-binding globulin, BMI, vitamin B12 as the most important variables.

Discussion

The main goal of the present study was to identify significant determinants of early and late-onset menopause using data mining algorithms. These models will be able to effectively screen women who carry a higher probability of early or late-onset menopause. This is significant because the unusual timing of menopause may indicate

Table 1: Socio-demographic characteristics of naturally menopause women, stratified by age at natural menopause, NHANES, 2013-2014 (*n*=762)

	Age at menopause *n* (%)			*P*
	Early (*n*=132)	Timely (*n*=575)	Late-onset (*n*=55)	
Race				
Mexican-American	(18.2) 24	75 (13)	3 (5.5)	0.325
Other Hispanic	15 (11.4)	61 (10.6)	8 (14.5)	
Non-Hispanic white	58 (43.9)	275 (47.8)	30 (45.5)	
Non-Hispanic black	21 (15.9)	85 (14.8)	10 (18.2)	
Other	14 (10.6)	79 (13.7)	4 (7.3)	
Marital status				
Married	(41.7) 55	286 (49.7)	(41.8) 23	0.388
Widowed	(20.5) 27	116 (20.2)	(25.5) 14	
Divorced	(18.9) 25	108 (18.8)	(18.2) 10	
Separated	8 (6.1)	12 (2.1)	(3.6) 2	
Never married	13 (9.8)	42 (7.3)	(7.3) 4	
Living with a partner	4 (3)	11 (1.9)	2 (3.6)	
Education level				
<9th grade	24 (18.2)	64 (11.1)	5 (9.1)	0.183
9-11th grade	18 (13.6)	66 (11.5)	7 (12.7)	
High school graduate	36 (27.3)	132 (23)	11 (20)	
Some college	33 (25)	168 (29.3)	18 (32.7)	
College graduate or above	21 (15.9)	144 (25.1)	14 (25.5)	
		Mean±SD*		
Age at study time (years)	62.8±10.6	64.1±9.1	67.4±6.3	0.009
Years since final menstrual periods	23.1±11.5	14.1±9.2	9.4±5.9	0.000
Total number of people in the Household	2.9±1.7	2.6±1.5	2.4±1.4	0.096
Ratio of family income to poverty	2.1±1.5	2.5±1.5	2.8±1.7	0.019

*Standard deviation

not only the loss of fertility but also an increased risk for various mid-life diseases and problems. Many of these diseases can be prevented by timely intervention, through lifestyle modification. The important contribution of the present work is that we searched the NHANES, a large population-based survey, for menopause timing determinant factors via data mining analytical approach. Menarche age and BMI are among the important contributors of early and late-onset menopause. Models trained using RF outperformed LR and SVM for ANM classification. Data mining has generated hypotheses that MMA, vitamin B12, SHBG, and vitamin D are possibly correlated to menopause timing.

The RF models developed in the study surpass LR and SVM. As suggested by a large body of literature RF outperforms SVM, however, the opposite has been reported too.[14] We are aware of no studies that have classified early or late-onset menopause using data mining approaches. Therefore, it is not possible to compare the current study with similar studies. The consistency of the performance metrics across imputed datasets suggests that imputation has been produced nearly similar data.

Our findings are consistent with previous studies that have established the association between age at menarche and ANM.[15] It is not completely clear whether early menarche cause early or late-onset menopause. The overall evidence is mixed. No correlation between the age of menarche and the age of menopause was reported in some studies.[16-18] A pooled analysis of nearly 50,000 postmenopausal women from nine observational studies in the UK, Scandinavia, Australia, and Japan, concluded that the risk of premature and early menopause increased by 80% for women with early menarche.[15]

In the present study, we found that MMA and vitamin B12 highly contribute to early and late-onset menopause classification. MMA is a carrier of vitamin B12, which is necessary for human metabolism and energy production, and its level is a biomarker for vitamin B12 deficiency.[19] Serum vitamin B12 concentrations are frequently low in the elderly.[20,21] Previous studies reported that lack of estrogen (menopause) affects the requirements for the B vitamins, including B12, for maintaining low blood homocysteine concentrations,[22] however, no study has explored the effect of vitamin B12 on menopause timing. Therefore, there is no clear explanation for the

Table 2: Distribution of features included in data mining algorithms, stratified by age at natural menopause, NHANES, 2013-2014

	Age at natural menopause (*n*=762) *n* (%)			*P*
	Early (*n*=132)	Timely (*n*=575)	Late-onset (*n*=55)	
Race				
Mexican-American	(18.2) 24	75 (13)	3 (5.5)	0.325
Other Hispanic	15 (11.4)	61 (10.6)	8 (14.5)	
Non-Hispanic white	58 (43.9)	275 (47.8)	30 (45.5)	
Non-Hispanic black	21 (15.9)	85 (14.8)	10 (18.2)	
Other	14 (10.6)	79 (13.7)	4 (7.3)	
Marital status				
Married	(41.7) 55	286 (49.7)	(41.8) 23	0.388
Widowed	(20.5) 27	116 (20.2)	(25.5) 14	
Divorced	(18.9) 25	108 (18.8)	(18.2) 10	
Separated	8 (6.1)	12 (2.1)	(3.6) 2	
Never married	13 (9.8)	42 (7.3)	(7.3) 4	
Living with a partner	4 (3)	11 (1.9)	2 (3.6)	
Education level				
<9th grade	24 (18.2)	64 (11.1)	5 (9.1)	0.183
9-11th grade	18 (13.6)	66 (11.5)	7 (12.7)	
High school graduate	36 (27.3)	132 (23)	11 (20)	
Some college	33 (25)	168 (29.3)	18 (32.7)	
College graduate or above	21 (15.9)	144 (25.1)	14 (25.5)	
History of prior pregnancy (yes)	117 (88.6)	521 (90.9)	49 (89.1)	0.683
History of diabetes in pregnancy (yes)	7 (6)	29 (5.6)	1 (2)	0.761
Had any babies weigh 9 lbs or more? (yes)	17 (15.3)	86 (17.2)	8 (17)	0.894
Birth control pills consumption (yes)	79 (60.3)	351 (61.3)	40 (72.7)	0.226
Hepatitis B core antibody (positive)	14 (11)	62 (11.2)	4 (7.4)	0.699
History of Cocaine/heroin/methamphetamine use (yes)	12 (13.3)	55 (13.8)	4 (12.1)	0.950
History of asthma (yes)	31 (23.5)	82 (14.3)	13 (23.6)	0.012
History of arthritis (yes)	69 (52.3)	260 (45.4)	39 (70.9)	0.001
History of thyroid problem (yes)	24 (18.2)	130 (22.6)	12 (21.8)	0.539
History of chronic obstructive pulmonary disease (yes)	13 (9.8)	21 (3.7)	1 (1.8)	0.005
Do vigorous recreational activities in a typical week (yes)	11 (8.3)	70 (12.5)	5 (9.1)	0.393
Do moderate recreational activities in a typical week (yes)	42 (31.8)	242 (42.1)	27 (49.1)	0.041
Used products in home to control insects (yes)	26 (19.7)	60 (10.4)	11 (20)	0.004
Used products to kill weeds (yes)	4 (3.1)	29 (5.2)	5 (9.3)	0.224
Smoke at least 100 cigarettes in life (yes)	61 (46.2)	214 (37.2)	21 (38.2)	0.160
History of vaginal, anal, or oral sex (yes)	85 (95.5)	377 (95.7)	34 (100)	0.460
		Mean±SD*		
Total number of people in the Household	2.9±1.7	2.6±1.5	2.4±1.4	0.096
Ratio of family income to poverty	2.1±1.5	2.5±1.5	2.8±1.7	0.019
Age at menarche	12.8±2.2	12.9±1.7	13±2.3	0.911
Number of pregnancies	3.9±2.2	3.6±2	3.4±1.9	0.299
Number of vaginal deliveries	2.7±2.2	2.6±2	2.5±1.8	0.854
Number of deliveries live birth result	3.2±1.9	2.9±1.8	2.8±1.7	0.196
Age at first live birth	22.7±4.7	22.7±4.7	22.7±4.8	0.691
Age at last live birth	30.5±6	30.9±5.9	29.8±6.1	0.480
Body Mass Index (kg/m^2)	30.2±7.8	28.9±7.1	30.6±7.6	0.078
Waist Circumference (cm)	101±16.1	97.9±15.2	101.7±16.2	0.047
Glycohemoglobin (%)	5.9±1.1	6.1±1.3	5.9±1.3	0.467
Methylmalonic Acid (nmol/L)	232.2±15.9	199.1±15.4	204.7±20.8	0.072
Testosterone (ng/dL)	18.9±1.4	22.1±1.4	23.4±1.5	0.167

Contd...

Table 2: Contd...

	Age at natural menopause (*n*=762) *n* (%)			*P*
	Early (*n*=132)	Timely (*n*=575)	Late-onset (*n*=55)	
Estradiol (pg/mL)	7.2±1.7	8.6±1.6	7.5±0.3	0.578
Sex hormone binding globulin (nmol/L)	78.8±4.1	70.5±4.5	78±4.7	0.097
25OHD2 + 25OHD3 (nmol/L)	77.4±3.4	78.2±3.7	74.9±3.4	0.773
Vitamin B12 (pg/mL)	681.6±62.7	757.6±68	689.1±61	0.570
Number of alcohol drinks over the past 12 months	2.8±0.5	3±0.5	2.5±0.4	0.509
Age at first sex	18.3±4.4	19±4.4	18.6±3.4	0.409

*Standard deviation

Table 3: The performance of the Data mining models for the classification of women age at menopause using full set of features

ANM*	RF† balanced				SVM‡ balanced				LR§ balanced			
	Sen‖	Spe¶	Brier score	AUROC**	Sen	Spe	Brier score	AUROC	Sen	Spe	Brier score	AUROC
Early vs. timely												
imputed data #1	94.3	96.4	0.072	98.1	63.8	67.1	0.231	65.3	61.8	64.7	0.214	70.5
imputed data #2	91.7	95.7	0.073	98.2	68.4	70.9	0.229	69.6	65.9	68.7	0.219	72.5
imputed data #3	95.5	95.9	0.071	98.4	66	66.2	0.225	66.1	65.9	68.7	0.212	73.2
imputed data #4	94.3	97	0.069	98.4	64.1	66	0.238	65	66.4	65.9	0.213	71.5
imputed data #5	93.7	95.8	0.068	98.1	65.2	68.5	0.237	66.7	72.2	64.7	0.217	73.6
combined estimate	96.9	97.6	0.051	99.0	67.7	69.6	0.211	68.6	67.3	68.4	0.200	75.7
late-onset vs. timely												
imputed data #1	98.8	99	0.019	99	70.8	92.9	0.231	78.03	89	65.9	0.210	79.5
imputed data #2	95.6	99	0.019	99	69.8	84.2	0.228	75.1	87.8	65.9	0.209	78.3
imputed data #3	99.4	99	0.018	99	69.8	85.4	0.227	75.4	84.9	67	0.207	77/7
imputed data #4	99.4	99	0.020	99	71.6	90	0.238	78	86.1	65.3	0.205	76.1
imputed data #5	99.4	99	0.018	99	69.6	91	0.239	76.5	86.7	65.9	0.211	77/7
combined estimate	99.5	99.8	0.005	99	74.4	85.5	0.200	78.9	82.3	74.5	0.119	84.1

*Age at natural menopause; †Random forest; ‡Support vector machine; §Logistic regression; ‖Sensitivity; ¶Specificity; **Area Under Receiver Operating Characteristic

observed relationship. Future research is essential to examine the possibility of an association between MMA, vitamin B12, and ANM.

SHBG was also found to be an important contributor to the unusual timing of menopause. SHBG binds to three sex hormones, including testosterone, dihydrotestosterone, and estradiol to regulate these hormone levels in the body. Evidence on the link between SHBG and ANM is lacking. The age-related trend of SHBG level among women is not clear and can be affected by many factors such as BMI and fasting insulin.[23] A meta-analysis of data retrieved from nine studies that investigated serum androgen profiles in women with premature ovarian failure found that these women did not seem to have a statistically significant difference compared to fertile women with regards to SHBG levels.[24]

The impact of vitamin D on female reproduction and the related disease has been thoroughly researched, however, the evidence on the link between vitamin D and ANM is scant. A prospective study reported that a higher level intake of vitamin D decreases the risk of early menopause.[25] Our work identified vitamin D as an important determinant of early menopause. This association may be biologically explainable. Active metabolites of vitamin D regulate genes involved in estrogen synthesis.[26] Also, follicle-stimulating

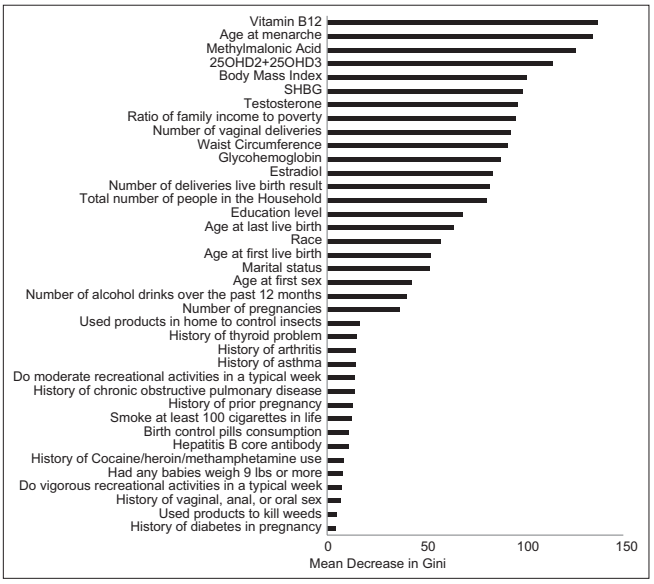

Figure 1: Features importance for early menopause based on mean decrease in Gini within RF model

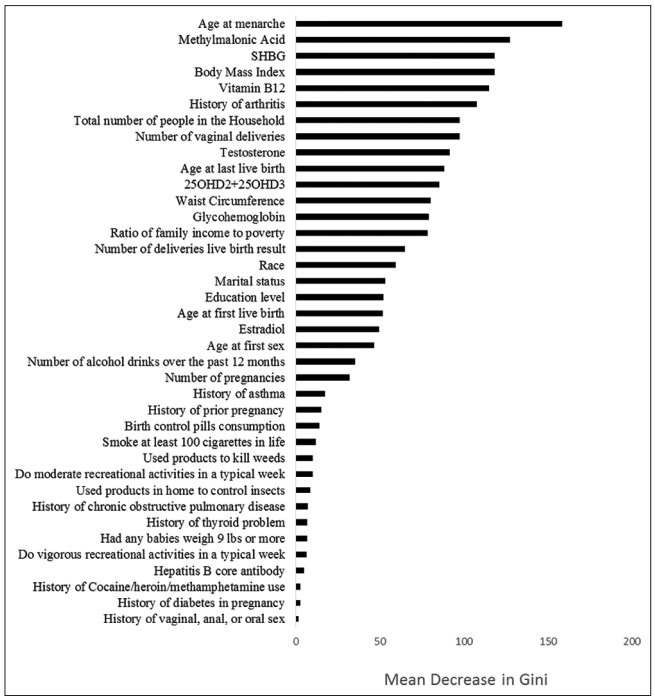

Figure 2: Features importance for late-onset menopause based on mean decrease in Gini within RF model

hormone (FSH) was reported to be inversely associated with vitamin D, and FSH is a biomarker of ovarian reserve, which rises across the late reproductive lifespan.[27]

The present study found BMI as an important variable in the classification of menopause age on both early and late-onset menopause. This finding is in line with existing evidence. Several previous studies documented an association between BMI and ANM. The higher the BMI, the later the age at menopause. BMI is the major determinant of endogenous estrogen level, therefore women with lower BMI are at risk of early menopause,[28-30] and those with higher BMI are supposed to have higher levels of estradiol and estrone in the body, and consequently later ANM.[31]

It is important to be aware of the limitations of cross-sectional studies. Menopause is a condition, with extensive physiological and psychological changes, in combination with the advancement of age. One limitation of the cross-sectional study design is that because the exposure and outcome are simultaneously assessed, there is generally no evidence of a temporal relationship between explored variables and ANM. Women were menopause when they participated in the survey and some of their health information is related to their current situation. Some of the factors examined in that age range, including the serum level of biological markers may not reflect the status of these factors in whole life or the years before and around menopause. Future research should employ longitudinal designs to validate cross-sectional findings obtained in the present study to ascertain the temporal trends for predicting ANM.

This paper intended to find correlated factors of early and late-onset menopause using three popular data mining approaches. The RF models were consistently better classifiers than other models. Age at menarche and BMI have a contributing effect on menopause timing. Future research focusing on the effect of the level of vitamin D, vitamin B12, and SHBG on menopause timing is proposed and will produce invaluable information for better prediction of the age at which menopause starts.

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

Received: 27 Oct 20 **Accepted:** 21 Jan 21
Published: 30 Dec 21

References

1. Amanvermez R, Tosun M. An update on ovarian aging and ovarian reserve tests. Int J Fertil Steril 2016;9:411-5.
2. Research on the menopause in the 1990s. Report of a WHO Scientific Group. World Health Organization technical report series. 1996;866:1-107.
3. O'Connor KA, Holman DJ, Wood JW. Menstrual cycle variability and the perimenopause. Am J Hum Biol 2001;13:465-78.
4. Cagnacci A, Pansini FS, Bacchi-Modena A, Giulini N, Mollica G, De Aloysio D, *et al*. Season of birth influences the timing of menopause. Hum Reprod 2005;20:2190-3.
5. Pikwer M, Bergstrom U, Nilsson JA, Jacobsson L, Turesson C. Early menopause is an independent predictor of rheumatoid arthritis. Ann Rheum Dis 2012;71:378-81.
6. Hong JS, Yi SW, Kang HC, Jee SH, Kang HG, Bayasgalan G, *et al*. Age at menopause and cause-specific mortality in South Korean women: Kangwha Cohort Study. Maturitas 2007;56:411-9.
7. Jacobsen BK, Heuch I, Kvåle G. Age at natural menopause and all-cause mortality: A 37-year follow-up of 19,731 Norwegian women. Am J Epidemiol 2003;157:923-9.
8. Ruth KS, Murray A. Lessons from genome-wide association studies in reproductive medicine: Menopause. Semin Reprod Med 2016;34:215-23.
9. Ahmed K, Jahan P, Nadia I, Ahmed F, Abdullah Al E. Assessment of menopausal symptoms among early and late menopausal midlife Bangladeshi women and their impact on the quality of life. J Menopausal Med 2016;22:39-46.
10. Kelsey TW, Anderson RA, Wright P, Nelson SM, Wallace WH. Data-driven assessment of the human ovarian reserve. Mol Hum Reprod 2012;18:79-87.
11. Malinowski J, Farber-Eger E, Crawford DC. Development of a data-mining algorithm to identify ages at reproductive milestones in electronic medical records. Pac Symp Biocomput 2014:376-87.
12. Faubion SS, Kuhle CL, Shuster LT, Rocca WA. Long-term health consequences of premature or early menopause and considerations for management. Climacteric 2015;18:483-91.
13. Canonico M, Plu-Bureau G, O'Sullivan MJ, Stefanick ML, Cochrane B, Scarabin PY, *et al*. Age at menopause, reproductive

history and venous thromboembolism risk among postmenopausal women: The women's health initiative hormone therapy clinical trials. Menopause 2014;21:214-20.
14. Statnikov A, Aliferis CF. Are random forests better than support vector machines for microarray-based cancer classification? AMIA Annu Symp Proc 2007;2007:686-90.
15. Mishra GD, Pandeya N, Dobson AJ, Chung HF, Anderson D, Kuh D, *et al*. Early menarche, nulliparity and the risk for premature and early natural menopause. Hum Reprod 2017;32:679-86.
16. Rizvanovic M, Balic D, Begic Z, Babovic A, Bogadanovic G, Kameric L. Parity and menarche as risk factors of time of menopause occurrence. Med Arch 2013;67:336-8.
17. Zsakai A, Mascie-Taylor N, Bodzsar EB. Relationship between some indicators of reproductive history, body fatness and the menopausal transition in Hungarian women. J Physiol Anthropol 2015;34:35.
18. Bjelland EK, Hofvind S, Byberg L, Eskild A. The relation of age at menarche with age at natural menopause: A population study of 336 788 women in Norway. Hum Reprod 2018;33:1149-57.
19. Hannibal L, Lysne V, Bjorke-Monsen AL, Behringer S, Grunert SC, Spiekerkoetter U, *et al*. Biomarkers and algorithms for the diagnosis of vitamin B12 deficiency. Front Mol Biosci 2016;3:27.
20. Allen LH. Vitamin B-12. Adv Nutr 2012;3:54-5.
21. Carmel R, Howard JM, Green R, Jacobsen DW, Azen C. Hormone replacement therapy and cobalamin status in elderly women. Am J Clin Nutr 1996;64:856-9.
22. Allen LH, Miller JW, de Groot L, Rosenberg IH, Smith AD, Refsum H, *et al*. Biomarkers of Nutrition for Development (BOND): Vitamin B-12 review. J Nutr 2018;148:1995S-2027S.
23. Maggio M, Lauretani F, Basaria S, Ceda GP, Bandinelli S, Metter EJ, *et al*. Sex hormone binding globulin levels across the adult lifespan in women--The role of body mass index and fasting insulin. J Endocrinol Invest 2008;31:597-601.
24. Soman M, Huang LC, Cai WH, Xu JB, Chen JY, He RK, *et al*. Serum androgen profiles in women with premature ovarian insufficiency: A systematic review and meta-analysis. Menopause 2019;26:78-93.
25. Purdue-Smithe AC, Whitcomb BW, Szegda KL, Boutot ME, Manson JE, Hankinson SE, *et al*. Vitamin D and calcium intake and risk of early menopause. Am J Clin Nutr 2017;105:1493-501.
26. Hong SH, Lee JE, An SM, Shin YY, Hwang DY, Yang SY, *et al*. Effect of vitamin D3 on biosynthesis of estrogen in porcine granulosa cells via modulation of steroidogenic enzymes. Toxicol Res 2017;33:49-54.
27. Jukic AMZ, Steiner AZ, Baird DD. Association between serum 25-hydroxyvitamin D and ovarian reserve in premenopausal women. Menopause (New York, NY) 2015;22:312-6.
28. Ahuja M. Age of menopause and determinants of menopause age: A PAN India survey by IMS. J Midlife Health 2016;7:126-31.
29. Akahoshi M, Soda M, Nakashima E, Tominaga T, Ichimaru S, Seto S, *et al*. The effects of body mass index on age at menopause. Int J Obes Relat Metab Disord 2002;26:961-8.
30. Tao X, Jiang A, Yin L, Li Y, Tao F, Hu H. Body mass index and age at natural menopause: A meta-analysis. Menopause 2015;22:469-74.
31. McTiernan A, Wu L, Chen C, Chlebowski R, Mossavar-Rahmani Y, Modugno F, *et al*. Relation of BMI and physical activity to sex hormones in postmenopausal women. Obesity (Silver Spring) 2006;14:1662-77.