Important factors related to self-perceived health among older men: a machine learning approach.

SUPPLEMENTAL MATERIAL CONTENT

Supplementary text - Data handling and the derivation of new variables

Health conditions

The following health conditions diagnosed by medical practitioners were self-reported by the participants: abdominal aortic aneurysm (AAA); angina; atrial fibrillation; coronary artery bypass grafting (CABG); carotid artery stenosis; heart failure; myocardial infarction; stroke; valvular heart disease; asthma; chronic obstructive pulmonary disease (COPD); sleep apnea; tuberculosis; and other lung disease. The following reported health conditions were kept as dichotomous variables: cancer; diabetes mellitus; hyperlipidemia; hypertension; and rheumatologic disease.

Self-reported factors

The participants also self-reported height, weight, highest educational level (elementary, upper secondary school, professional school, or university), smoking status (everyday, sometimes, former, or never), number of cigarettes per day, duration of smoking, exercise frequency (every day, three to six times a week, one to three times a week, less than once a week), sleep quality (usually slept: very well, well, quite well, bad, very bad), and average sleep duration (4 or less hours, 5 hours, 6 hours, 7 hours, 8 hours, 9 hours, or 10 hours or more).

Derivation of new variables

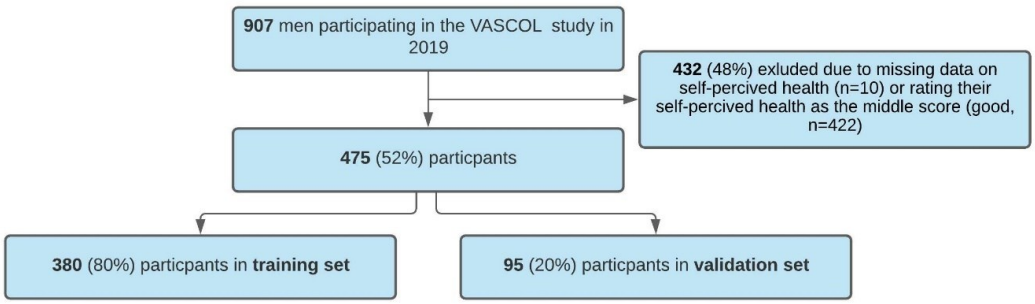
Body mass index (BMI) was calculated as kg/m^2. Pack-years were calculated based on the number of years of smoking and the number of cigarettes per day. Education level was dichotomized as either university degree or no university degree, and smoking status was dichotomized as ever smoker or never smoker. **Cardiovascular** (AAA, angina, atrial fibrillation, coronary artery bypass grafting (CABG), carotid artery stenosis, heart failure, myocardial infarction, stroke, valvular heart disease) and **respiratory diseases** (asthma, chronic obstructive pulmonary disease (COPD), sleep apnea, tuberculosis, other lung disease) were categorized as dichotomous variables to compensate for the expected low prevalence of the individual health conditions. The other health conditions (diabetes, hyperlipidemia, hypertension, and rheumatologic disease) were kept as individual dichotomous variables. The variables that were used to derive new variables, such as height and weight, were excluded from the dataset.

Table S1 – Characteristics of 475 men in the training and validation sets.

		Self-perceived health	
Factors	**Training set**	**Validation set**	**P value**
	380 (80%)	95 (20%)	
Worse self-perceived health	221 (58.1)	61 (64.2)	0.338
BMI	27.12 (4.18)	26.89 (3.46)	0.622
University degree	82 (21.6)	14 (14.7)	0.179
Ever smoker	255 (67.1)	58 (61.1)	0.321
Pack years of smoking	12.4 (14.9)	18.1 (30.5)	0.149
Exercise frequency			0.679
Less than once a week	63 (16.6)	19 (20.0)	
1-3 times a week	105 (27.6)	29 (30.5)	
3-6 times a week	121 (31.8)	25 (26.3)	
Everyday	91 (23.9)	22 (23.2)	
Standard units of alcohol	6.61 (6.35)	5.74 (6.10)	0.233
Sleep quality			0.780
Very bad	4 (1.1)	1 (1.1)	
Bad	55 (14.5)	10 (10.5)	
Quite good	123 (32.4)	36 (37.9)	
Good	109 (28.7)	28 (29.5)	
Very good	89 (23.4)	20 (21.1)	
Sleep duration			0.258
4 hours or less	7 (1.8)	4 (4.2)	
5 hours	37 (9.7)	6 (6.3)	
6 hours	57 (15.0)	20 (21.1)	
7 hours	148 (38.9)	29 (30.5)	
8 hours	104 (27.4)	26 (27.4)	
9 hours	24 (6.3)	8 (8.4)	
10 hours or more	3 (0.8)	2 (2.1)	
Symptoms			
Anxiety	1.03 (2.02)	1.03 (1.85)	1.000
Appetite	0.74 (1.68)	0.58 (1.77)	0.410
Breathlessness	2.18 (2.66)	1.69 (2.34)	0.107
Depression	1.51 (2.22)	1.37 (2.13)	0.588
Drowsiness	2.39 (2.43)	2.19 (2.26)	0.461
Fatigue	2.82 (2.60)	2.53 (2.37)	0.323
Nausea	0.63 (1.50)	0.48 (1.30)	0.397
Pain	2.81 (2.62)	2.96 (2.58)	0.623
Health conditions			
Cancer	75 (19.7)	14 (14.7)	0.332
Cardiovascular disease	146 (38.4)	33 (34.7)	0.586
Diabetes	61 (16.1)	18 (18.9)	0.600
Hypertension	221 (58.2)	54 (56.8)	0.908
Hyperlipidemia	99 (26.1)	28 (29.5)	0.586
Respiratory disease	70 (18.4)	15 (15.8)	0.654
Rheumatologic disease	24 (6.3)	6 (6.3)	1.000

All values are presented as the mean (standard deviation) or frequency (%).

Figure S1- Flowchart of exclusion of participants and training/validation split.

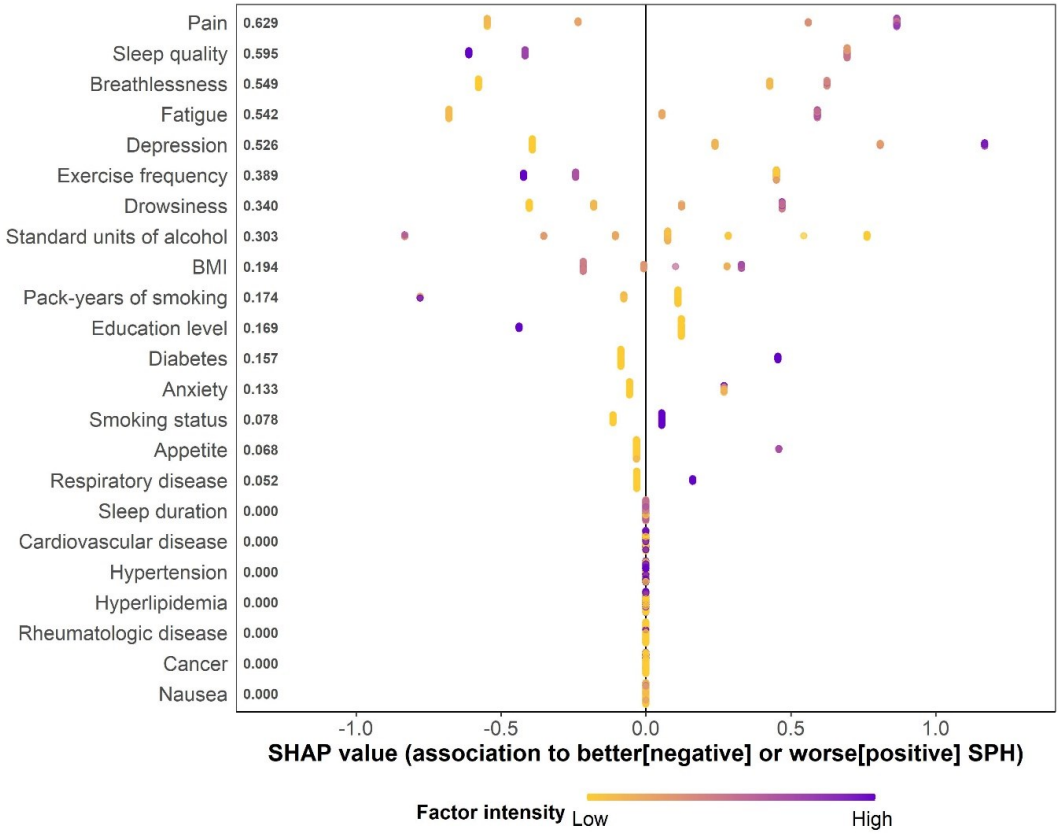


Self-perceived health was measured with the first item of the SF-12v2: "*In general, would you say your health is:*", and the participants were categorized as having either *better* (answers: excellent or very good), *middle* (answer: good), or *worse* (answers: fairly, or poor) self-perceived health. The *middle* category was removed from the dataset to differentiate the outcome categories from each other. A total of 221 (58%) participants in the training set and 61 (64%) in the validation set perceived their health as worse.

Table S2 – Hyperparameters of the final XGBoost model

Hyperparameter in the XGBoost R-package	**Explanation**	**Value**
gamma	Minimum loss reduction to create a new leaf node of the tree	0
eta	Learning rate	0.3
max_depth	Maximum tree depth	0
colsample_bytree	Subsample ratio for each tree	0.8
min_child_weight	Minimum sum of instance weight to continue building process	1

Figure S2 – Individual SHAP values by factor intensity.



Dots represent individual participants, and participants are stacked vertically to indicate that multiple participants have the same SHapley Additive exPlanations (SHAP) values. The SHAP value (x-axis) represents the strength and direction of the association for worse (positive SHAP) or better (negative SHAP) self-perceived health for the individual participants in the validation set. The *factor intensity* represents the scaled variable values from low to high among the individual participants and is presented as color intensity. SHAP absolute mean values are presented to the right of the factor names. A higher SHAP absolute mean value corresponds to greater importance for predicting self-perceived health as *better* or *worse* among all participants in the validation set. Note: A higher appetite feature value corresponds to a worse appetite. A higher feature value of sleep quality corresponds to better sleep quality. The presence of a condition, being ever smoker (smoking status) and having a university degree (education level) are marked with purple. Abbreviations: SPH = self-perceived health; BMI = body mass index.