This is the post print version of the article, which has been published in Scandinavian journal of public health. 2018, 46 (5), 557-564 https://doi.org/10.1177/1403494817736944.

Suitability of random forest for epidemiological research: exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design

Noora Kanerva[1,2], PhD, Jukka Kontto[2], MSc, Maijaliisa Erkkola[3], PhD, Jaakko Nevalainen[4] PhD, Satu Männistö[2], PhD.

[1] Department of Public Health, University of Helsinki, Helsinki, Finland

[2] Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland

[3] Nutrition Unit, University of Helsinki, Helsinki, Finland

[4] School of Health Sciences, University of Tampere, Tampere, Finland

Correspondence: Noora Kanerva, PhD; Department of Public Health, University of Helsinki, POB 20, 00014 University of Helsinki, Helsinki, Finland; E-mail: noora.kanerva@helsinki.fi; Phone: +358408272356.

Authorship: SM, MLE and JN were responsible for the original study idea and analysis plan. NK and JK conducted the statistical analyses. NK wrote the manuscript and had

primary responsibility of the final content. All coauthors have critically reviewed and approved the manuscript.

Short title: Random forest analysis in epidemiological research

Word count: 3248

Abstract

Aims: Factors that contribute to development of overweight are numerous and form a complex structure with many unknown interactions and associations. We aimed to explore this structure, i.e. the mutual importance or hierarchy of sociodemographic and lifestyle-related risk factors of overweight, using a machine-learning technique called random forest. The results were compared with traditional logistic regression analysis.

Methods: The cross-sectional FINRISK 2007 Study included 4 757 Finns (25–74 years). Information on participants' lifestyle and sociodemographic characteristics were collected with questionnaires. Diet was assessed, using a validated food-frequency questionnaire. Height and weight were measured. Participants with BMI ≥25 kg/m^2 were classified as overweight. R-statistical software was used to run random forest analysis (RF) ('randomForest') to derive estimates for variable importance and out-of-bag error, which were compared to a logistic regression model (LR).

Results: In total, 704 (32%) men and 1 119 (44%) women had normal BMI, whereas 1 502 (69%) men and 1 432 (57%) women had BMI ≥25. Estimated error rates for the models were similar (RF vs LR: 42% vs. 40% for men, 38% vs. 35% for women). Both models ranked

age, education and physical activity as the most important risk factors for overweight, but RF ranked macronutrients (carbohydrates and protein) more important compared to LR. Conclusion: RF did not demonstrate higher power in variable selection compared to LR in our study. The features of RF are more likely to appear beneficial in settings with a larger number of predictors.

Introduction

Overweight (body mass index [BMI] ≥25.0 kg/m^2) and obesity (BMI ≥30.0 kg/m^2) are worldwide health problems that have been ranked as the sixth most important factor contributing to mortality and morbidity from chronic diseases [1]. Large meta-analyses have reported substantial increases in BMI over the last 30 to 40 years [2]. The number of overweight men increased from 29% to 37% whereas among women it increased from 30% to 38% [3]. Thus, effective public health approaches to stop people from gaining extra weight are urgently needed.

Overweight is fundamentally the result of an imbalance between intake and expenditure of energy. Factors that contribute to this imbalance are numerous: socioeconomic status, smoking, inactive lifestyle, unhealthy eating habits, mental health, living environment, insufficient sleep, and genetic factors are known to affect the development of overweight [4]. These risk factors are manifested in various combinations in individuals. For example, for some, overweight may be a symptom of psychological problems such as depression or insufficient sleep, and for others it could be solely the result of an inactive lifestyle or genetic

susceptibility. The inability to recognize and respond to these various combinations at population level creates a barrier for effective prevention and treatment policy [5, 6].

During the last 10 years, a machine-learning technique called random forest has been developed to provide a solution for classification problems [7]. The advantage of such a technique is that it takes into account each predictor individually, even if the association with the outcome is not linear, and the multivariate interactions with the other predictors. For genetic and other high-dimensional data, random forest analysis has shown great potential for classification and ranking of relevant variables [8]. We wanted to explore whether random forest analysis would also prove beneficial in other areas of epidemiology in which the number of variables is considerably lower compared to genetic data, but where those variables also form complex structures that involve many unknown interactions and associations. As a case study, we decided to explore the mutual importance (i.e. hierarchy) of various sociodemographic and lifestyle-related risk factors of overweight. The results obtained using random forest were compared to a traditional statistical analysis.

Study population and methods

In this study, we used a population-based sample of men and women ages 25 to 74 years who participated in two phases of the National FINRISK 2007 Study. Between January and March 2007, a random sample of 10 000 participants was drawn from the Finnish Population Information System in five large geographical areas [9]. The sample was stratified by sex, 10-year age groups, and area. The participants were sent an invitation letter to a health examination with a self-administrative health questionnaire. Of the invited subjects, 6 258 participated in the health examination (participation rate of 63%).

The second study phase that aimed to gather more precise information on obesity was conducted between April and June 2007 [10]. This phase included a detailed health examination and several questionnaires. Of the subjects who participated in the first phase, 5 024 attended to the second phase (participation rate of 80%). For this study's purposes, we excluded participants with a missing or incomplete food frequency questionnaire (FFQ), those with no anthropometric data available, and women who were pregnant, which left us with 2 206 (44% of those who participated) men and 2 551 (51% of those who participated) women for the analyses.

The National FINRISK 2007 Study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving human subjects were approved by the Ethics Committee of the Hospital District of Helsinki and Uusimaa. Written informed consent was obtained from all participants.

Information on participants' age, sex and living area were obtained from the Population Information System. The health questionnaires assessed participants' education, smoking status and leisure-time physical activity (PA). Participants' education was assessed in number of years spent in education. Smoking was assessed as the number of daily smoked cigarettes. The level of PA was assessed as activities outside of work using four categories: inactive (mainly light activities, e.g. reading, watching television), moderately active (e.g. walking, cycling or gardening at least 4 h per week), highly active (physically demanding activities, e.g. running, cross-country skiing or swimming at least 3 h per week), and extremely active (competition sports–related exercise several times per week). All participants were asked to report how many hours they usually sleep during night-time. Furthermore, women who had children were asked to report the number of times they gave birth on the questionnaire.

A validated and self-administrative FFQ was used to assess participants' habitual intake of 131 food items and mixed dishes [11, 12]. Participants filled in the FFQ at the study site, during the health examination of the second study phase. The subjects were asked to indicate the average consumption frequency of each FFQ item by using nine frequency categories ranging from 'never or seldom' to 'six or more times a day'. The predefined portion sizes appeared as household and natural units (e.g. glass, slice) on the FFQ and were fixed separately for both men and women based on information obtained from the National FINDIET 2007 Survey [13]. The participants were also able to report other frequently consumed foods not listed. A study nurse reviewed the FFQ after each participant filled it in. Data were entered into the study database and the average daily food, nutrient and energy intakes were calculated using the Finnish National Food Composition Database (Fineli®) and in-house software [14].

A trained study nurse measured participants' weight to the nearest 0.1 kg and height to the nearest 0.1 cm. Participants were allowed to wear only light clothing and no shoes during the anthropometric measurements. BMI was calculated by dividing weight (kg) by the square of height (m^2). All measurements were done according to standardized international

recommendations [15]. Participants with BMI <25.0 kg/m^2 were categorized as normal weight, whereas participants with BMI ≥25.0 kg/m^2 were categorized as overweight [16].

Data analysis

The data were analysed with R-statistical software version 3.0.2 [17]. Participants' characteristics are presented separately for men and women as median and 1st and 3rd quartiles or %. Spearman correlation coefficients between sociodemographic and lifestyle variables with 95% confidence intervals were calculated and illustrated in a correlogram. To avoid the confounding effect of total energy intake related to differences in PA and body nutrition, intakes were energy-adjusted using the Willett residual method [18], which takes into account the amount of total energy in relation to the intake of nutrients between participants.

Random forest analysis [7] is based on an ensemble of classification trees [19]. In a classification tree, a data set is split into two subgroups (nodes) using a value of a correlate, which maximizes the homogeneity of the subgroups. After the first split, the process is applied to each node recursively until the nodes reach a minimum size or until no improvement in the splitting can be made. Random forest is an ensemble of hundreds or

thousands of classification trees that are grown using a random subset of individuals and random selection of correlates. The out-of-bag (OOB) proportion of the data that is left outside the building of a tree is used as validation data to compute the classification error, which in the end is averaged over all trees. The difference between the OOB error resulting from a data set obtained through random permutation of the correlate of interest and the OOB error resulting from the original data set can be used as a measure of variable importance.

The possible weight subgroups were illustrated with a single classification tree ("rpart"). Random forest analysis ("randomForest") [7] was used to derive the classification and an estimate of exposure importance: the strength of association between weight of the correlates and outcome. For both men and women, 1 000 random subsets were drawn from the data to grow 1 000 classification trees. For dietary variables, we first ran separate random forest analyses for 65 foods and >100 nutrients. Of these models, due to the variable importance measure given by random forest and rationale from the public health perspective, total energy intake and macronutrients were selected in a combined model with other lifestyle and sociodemographic factors (age, living area, education, smoking status, PA, sleep duration and number of labours). In the combined model, a random selection of 5 correlates out of 14 correlates was sampled to derive each split in each tree. Node size and maximum number of

terminal nodes were not restricted. The OOB error and variable importance measures of the resulting random forest were compared with results obtained from traditional logistic regression analysis (glm-procedure in ‘base’).

Results

Participant characteristics are presented by sex and BMI in Table I. In total, 1 502 (69%) of men and 1 432 (57%) of women had BMI ≥25 kg/m^2. Overweight participants were older, had fewer educational years on average and had lower PA compared to normal-weight participants. Furthermore, fewer overweight men were never smokers, and they had higher energy intake compared to normal-weight men. Other lifestyle factors did not substantially differ between BMI classes. Correlations between the sociodemographic and lifestyle factors were mostly low (Spearman correlation coefficient r <0.20), except between the dietary variables (highest correlation between energy-adjusted carbohydrates and fat r = -0.75 for men and r = -0.84 for women) (Supplementary Figures I and II).

Examples of weight subgroups that share similar sociodemographic and lifestyle characteristics are illustrated as simple classification trees in Supplementary Figures III and

IV. In each branch of the classification tree, subgroups with more normal-weight participants are classified to the left, and subgroups with more overweight participants are classified to the right. As shown in Supplementary Figure III, men younger than 32.5 years with energy intake less than 10 042 kJ/d were most likely to have normal BMI, whereas men older than 32.5 years were most likely to be overweight. In women, those who were younger than 45.5 years with moderate or higher PA were most likely normal weight, whereas those older than 45.5 years with moderate or lower PA were most likely overweight (Supplementary Figure IV).

In random forest analysis, age was ranked as the most important factor in both men and women (Figure I). In men, intake of carbohydrates and alcohol, and education were the next most important factors (Figure Ia). In women, PA and education stood out as the second most important factors before intake of carbohydrates (Figure Ib). In the logistic regression analysis, age and PA were significantly associated with overweight in both men and women (Table II). Education was significantly associated with overweight only in women. None of the dietary factors associated statistically significantly with the odds of being overweight.

Estimated error rates for the random forest analysis compared to the logistic regression analysis were fairly similar, logistic regression having a slightly smaller error (OOB error estimate 41.6% vs. 39.7% for men; 37.7% vs. 35.1% for women) (Table III). If age, which was the strongest correlate for overweight, was entered into the model alone, the OOB error estimates attenuated in all models except men's random forest model (data not shown). In this model, the OOB error estimate improved, but this was due to increased sensitivity from 57.0% to 94.5% at the expense of specificity, which decreased from 61.5% to 12.5%. Thus, the random forest model for men actually attenuated when age was used as the only correlate in the model.

Discussion

The main contribution of this work is the exploration of whether random forest analysis would suit purposes of epidemiological research (other than genetic epidemiology) that has, thus far, relied heavily on traditional regression and survival analyses. As an example, we studied the mutual importance of sociodemographic and lifestyle factors that are known to be associated with overweight by using simple classification trees, random forest analysis

and logistic regression analysis. Perhaps surprisingly, the results between random forest analysis and logistic regression analysis were fairly similar. Both ranked age, education and PA as important factors, and also had quite similar classification accuracy. This could indicate that the associations between these variables and overweight are fairly linear and that the interactions between the studied variables are not major determinants of overweight.

The aetiology of overweight and its delayed complications are multifactorial, involving many behavioural factors, such as smoking, physical inactivity, unhealthy diet, high alcohol consumption, poor sleep and their social background factors [4]. Still, many studies tend to focus on a limited number of behaviours, even though it is obvious that human health should be studied as a whole [6]. During the last 10 years, many different patterning and summarizing tools have been implemented to gain a more holistic approach. For instance, latent class analysis was recently used to form profiles of PA and sleep behaviour, and then the association of these profiles to heart health was examined [20]. Furthermore, healthy dietary patterns have been examined in large-scale studies, using summary indices and principal component analysis instead of single foods and nutrients [21]. Applying machine-learning techniques to research is still very uncommon. Increasing numbers of studies are using tree-structured methodology in the field of obesity research [22, 23]. However, we did

not find any earlier published studies that applied random forest analysis in epidemiological research that would have involved the most basic determinants of health: PA, nutrition or sociodemographic factors. Our study, thus, is among the first to explore the suitability of using random forest analysis in examining the hierarchy of sociodemographic and lifestyle factors in this area.

Those research fields that have applied random forest have gained promising results. In studies related to clinical decision making, such as detecting patient groups with underlying diabetic retinopathy, classifiers based on random forest have had the highest prediction accuracy compared to other classification methods (e.g. logistic regression and support vector machine) [24, 25]. In genetic and other bioinformatics data analysis, random forest has outperformed standard statistical methods [26]. The method has also proven useful in studying risk factor dependencies, for instance, in road safety [27]. In our study, however, random forest had similar or even slightly poorer classification accuracy compared to traditional logistic regression analysis. Perhaps the low number of correlate variables and features of the variables, e.g. variability in their measurement scale (continuous variables) or number of classes (categorical variables) affected the accuracy. Moreover, the narrow range of variables probably contributed to nonsignificant findings between some variables (e.g.

smoking and sleeping) and overweight although these associations have been well established. For instance, >70% of the participants slept 7–8 hours per night and >80% of the participants were non-smokers. The weakness of random forest in comparison to structured models is the potential lack of interpretability of the effects of variables, which may limit understanding of what underlies the classification. Furthermore, repeatability of random forest analysis should be given more focus. It is known that prediction accuracy of single-decision trees varies considerably. This variation is reduced by summing up multiple trees so that each uses a randomly selected subset of individuals. Still the results of variable importance may vary because features of the variables may affect the variable selection in each node.

Despite the fact that random forest analysis did not show better classification power in our study, it has many beneficial features compared to the traditional methods. For instance, dietary factors are highly correlated, which limits their simultaneous exploration in logistic regression. However, random forest allows inclusion of such correlated data. Furthermore, interactions within dietary data, i.e. effect of a nutrient may depend on the level of intake of another nutrient (calcium absorption is dependent on vitamin D status), should be introduced to the logistic regression by the researcher. Random forest explores and finds these

interactions independently without any assumptions. Furthermore, single-decision trees give insight to possible subgroups—which combinations of lifestyle and sociodemographic factors lead to chronic diseases. Ideally, these subgroups may be introduced as new targets of public health actions.

Strengths of this study include a large population-based sample and the number of predictor variables available from several scientific branches, including nutrition, PA and sleep. BMI calculation was based on measured height and weight, and the international cut-off for overweight was used. Our study had some limitations too. First, the cross-sectional design of the study does not allow any assumptions on causality. Some variables that were based on self-report could be affected by misreporting, which may affect the results to some extent. For instance, overweight individuals are known to be prone to underestimation of their energy intake [28]. This systematic error may have led to attenuated correlation between energy intake and the risk of overweight, as well as decreased sensitivity and specificity of the model. Other commonly misreported foods are fruits and vegetables, sweet and fatty foods and alcoholic beverages, which are all known to associate with overweight [29]. The same problem of overestimation applies also to PA [30].

Random forest analysis may include some possible pitfalls. The selection of the tuning parameters of random forest may introduce subjectivity to the analysis. This possibility needs to be kept in mind when interpreting our results as measurement scale and number of categories varied in our predictors. Furthermore, correlation between predictors may in some cases induce confounding. Random forest produces variable importance lists regardless of whether the variables are informative or non-informative. Non-informative predictors that are highly correlated with other predictors tend to receive smaller importance measures than uncorrelated predictors. Thus, a non-informative predictor with a biased importance measure may outperform a moderately informative predictor.

In our cross-sectional study, random forest did not show any additional benefit compared to logistic regression that was conducted with a limited number of correlates. However, it has shown particular promise in settings where the number of predictors was closer to or above 100 [22–27]. Future studies should include prospective design and aim to include as many aspects of an individual as possible, such as sociodemographic and lifestyle predictors, metabolic and genetic predictors and psychological factors in random forest to gain a more holistic view of mutual dependencies of these factors in chronic diseases.

Declaration of conflicting interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study was funded by the Finnish Academy (Grant number 136895 and 263836 to SM).

References

[1] Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380(9859):2224-60.

[2] Finucane MM, Stevens GA, Cowan MJ, Danaei G, Lin JK, Paciorek CJ, et al. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. Lancet 2011;377(9765):557-67.

[3] Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and national prevalence of overweight and obesity in children and adults

during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2014;384(9945):766-81.

[4] Malik VS, Willett WC, Hu FB. Global obesity: trends, risk factors and policy implications. Nat Rev Endocrinol 2013;9(1):13-27.

[5] Crawford D, Ball K. Behavioural determinants of the obesity epidemic. Asia Pac J Clin Nutr 2002;11:S718-S721.

[6] Fuglestad PT, Jeffery RW, Sherwood NE. Lifestyle patterns associated with diet, physical activity, body mass index and amount of recent weight loss in a sample of successful weight losers. Int J Behav Nutr Phys Act 2012;9:79-5868-9-79.

[7] Breiman L. Random forests. Mach Learning 2001;45(1):5-32.

[8] Lunetta K, Hayward L, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004;5:32.

[9] Vartiainen E, Laatikainen T, Peltonen M, Juolevi A, Männistö S, Sundvall J, et al. Thirty-five-year trends in cardiovascular risk factors in Finland. Int J Epidemiol 2010;39(2):504-18.

[10] Konttinen H, Haukkala A, Sarlio-Lähteenkorva S, Silventoinen K, Jousilahti P. Eating styles, self-control and obesity indicators. The moderating role of obesity status and dieting history on restrained eating. Appetite 2009;53(1):131-4.

[11] Männistö S, Virtanen M, Mikkonen T, Pietinen P. Reproducibility and validity of a food frequency questionnaire in a case-control study on breast cancer. J Clin Epidemiol 1996;49(4):401-9.

[12] Kaartinen NE, Tapanainen H, Valsta LM, Similä ME, Reinivuo H, Korhonen T, et al. Relative validity of a FFQ in measuring carbohydrate fractions, dietary glycaemic index and load: exploring the effects of subject characteristics. Br J Nutr 2012;107(9):1367-75.

[13] Paturi M, Tapanainen H, Reinivuo H, Pietinen P. The National Findiet 2007 Survey (abstract in english). Publications of the National Public Health Institute B23/2008. Helsinki, 2008.

[14] Reinivuo H, Hirvonen T, Ovaskainen ML, Korhonen T, Valsta LM. Dietary survey methodology of FINDIET 2007 with a risk assessment perspective. Public Health Nutr 2010;13(6A):915-9.

[15] Tolonen H, Koponen P, Aromaa A, Conti S, Sidsel GI, Grotvedt L, et al. Recommendations for the health examination surveys in Europe. Publications of the National Public Health Institute B21/2008. Helsinki, 2008.

[16] World Health Organization. Obesity: Preventing and Managing the Global Epidemic. WHO technical Report Series 894. Geneva, 2000.

[17] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. Available at: http://www.R-project.org/.

[18] Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 1986;124(1):17-27.

[19] Breiman L, Friedman JH, Olshen R, Stone C. Classification and regression trees. Wadsworth; Belmont, CA, 1984.

[20] Wennman H, Kronholm E, Partonen T, Tolvanen A, Peltonen M, Vasankari T, et al. Physical activity and sleep profiles in Finnish men and women. BMC Public Health 2014;14:82-2458-14-82.

[21] Wirt A, Collins CE. Diet quality - what is it and does it matter? Public Health Nutr 2009;12(12):2473-92.

[22] Toschke AM, Beyerlein A, von Kries R. Children at High Risk for Overweight: A Classification and Regression Trees Analysis Approach. Obes Res 2005;13:1270-4.

[23] Roda C, Charreire H, Feuillet T, Mackenbach JD, Compernolle S, Glonti K, Bárdos H, Rutter H, McKee M, Brug J, De Bourdeaudhuij I, Lakerveld J, Oppert JM. Lifestyle correlates of overweight in adults: a hierarchical approach (the SPOTLIGHT project). Int J Behav Nutr Phys Act 2016;13:114.

[24] Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of random forests methods to diabetic retinopathy classification analyses. PLoS One 2014;9(6):e98587.

[25] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med 2016;44(2):368-74.

[26] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics 2012;99(6):323-9.

[27] Kwon OH, Rhee W, Yoon Y. Application of classification algorithms for analysis of road safety risk factor dependencies. Accid Anal Prev 2015;75:1-15.

[28] Stice E, Palmrose CA, Burger KS. Elevated BMI and male sex are associated with greater underreporting of caloric intake as assessed by doubly labeled water. J Nutr 2015;145:2412-8.

[29] Paalanen L, Männistö S, Virtanen MJ, Knekt P, Räsänen L, Montonen J, Pietinen P. Validity of a food frequency questionnaire varied by age and body mass index. J Clin Epidemiol 2006;59:994-1001.

[30] Pietiläinen KH, Korkeila M, Bogl LH, Westerterp KR, Yki-Järvinen H, Kaprio J, Rissanen A. Inaccuracies in food and physical activity diaries of obese subjects: complementary evidence from doubly labeled water and co-twin assessments. Int J Obes 2010;34:437-45.

Table I. Participant characteristics by gender and BMI in the National FINRISK 2007 Study.

	Men (n = 2206)						Women (n = 2551)					
	BMI < 25 kg/m^2 (n = 704)			BMI ≥ 25 kg/m^2 (n = 1502)			BMI < 25 kg/m^2 (n = 1119)			BMI ≥ 25 kg/m^2 (n = 1432)		
Characteristics	Median / %*	1st Q	3rd Q	Median / %	1st Q	3rd Q	Median / %	1st Q	3rd Q	Median / %	1st Q	3rd Q
Age, y	51	39	62	56	45	66	47	37	59	57	46	66
Number of labours	-	-	-	-	-	-	2	0	2	2	1	3
Living area, % :												
Helsinki/Vantaa	20			17			22			16		
Turku/Loimaa	18			21			22			22		
Northern Savo	19			20			19			22		
North Karelia	21			22			18			20		
Northern Ostrobothnia	21			21			19			19		
Educational years, y	13	10	16	11	9	15	14	11	17	12	9	15
Never smokers, %	52			42			66			67		
High physical activity, %	38			25			35			19		
Sleeping hours, h/night	7	7	8	7	7	8	7	7	8	7	7	8
BMI, kg/m^2	23.4	22.2	24.3	28.1	26.5	30.5	22.7	21.2	23.8	28.9	26.7	32.2
Energy intake, kJ/d	10710	8890	11440	11190	8970	13880	8930	7470	11020	8900	7050	11310
Carbohydrate, g/d †	277	255	300	272	249	296	288	265	310	287	264	308
Fibre, g/d †	27	21	32	27	22	32	32	26	38	32	27	39
Sucrose, g/d †	124	106	145	122	104	143	138	116	158	135	115	156
Fat, g/d †	84	75	91	83	75	92	81	73	90	81	72	89
Protein, g/d †	100	93	110	103	93	113	101	92	110	103	94	112
Alcohol, g/d †	16	5	32	18	5	36	8	2	18	6	<1	16

BMI: body mass index; Q: quartile; E%: intake as a percentage of the total energy intake.

* Values are given as percentages or median and the 1st and 3rd Quartiles.

† Nutrient intake has been energy-adjusted, using the Willett's residual method described in: Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 1986 Jul;124(1):17-27.

Table II. Association of sociodemographic and lifestyle factors with overweight or obesity in the National FINRISK 2007 Study: results from logistic regression analysis.

	Men		Women	
Predictor variables	β	*P*	β	*P*
Intercept	1.658	0.18	-1.363	0.88
Age, y	0.021	<0.001	0.033	<0.001
Number of labours*	-	-	0.114	<0.001
Area (ref. North Karelia)				
Northern Savo	0.076	0.61	0.004	0.97
Turku/Loimaa	0.038	0.80	-0.143	0.29
Helsinki/Vantaa (capital area)	-0.290	0.14	-0.442	<0.01
Northern Ostrobothnia	-0.066	0.65	-0.172	0.21
Education, y	-0.023	0.08	-0.031	<0.05
Smoking, cigarettes/d	-0.020	0.08	-0.017	0.18
PA (ref. Low PA)				
Moderate PA	-0.425	<0.01	-0.773	<0.001
High PA	-0.753	<0.001	-1.170	<0.001
Very high PA	-1.240	<0.001	-2.290	<0.01
Sleep, h/d	-0.068	0.14	-0.014	0.75
Energy, 1000 kJ/d	0.020	0.17	-0.001	0.24
Alcohol, g/d†	-0.013	0.34	0.003	0.81
Protein, g/d †	-0.015	0.49	0.015	0.37
Carbohydrates, g/d †	-0.032	0.13	-0.001	0.95
Fat, g/d †	-0.065	0.17	-0.003	0.93
Sucrose, g/d †	0.003	0.23	0.003	0.18
Fibre, g/d †	0.001	0.89	-0.003	0.62

PA: physical activity.

* Number of labours was added as predictive variable only when analysing data for women.

† Nutrient intake has been energy-adjusted, using the Willett's residual method described in: Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 1986 Jul;124(1):17-27.

Table III. Comparison of classification accuracy between random forest and logistic regression analysis in the National FINRISK 2007 Study.

	Statistical method	
	RF	LR
Men (*n* = 2206)		
n of true/false positives in participants with BMI ≥ 25 kg/m^2 (total *n* = 1502)	856 / 646	922/ 580
n of true/false negatives in participants with BMI < 25 kg/m^2 (total *n* = 704)	433 / 271	409 / 295
OOB error estimate, %	41.6	39.7
Sensitivity, %	57.0	61.4
Specificity, %	61.5	58.1
Women (*n* = 2551)		
n of true/false positives in participants with BMI ≥ 25 kg/m^2 (total *n* = 1432)	923 / 509	982/450
n of true/false negatives in participants with BMI < 25 kg/m^2 (total *n* = 1119)	667 /452	674/445
OOB error estimate, %	37.7	35.1
Sensitivity, %	64.5	68.6
Specificity, %	59.6	60.2

BMI: body mass index; LR: logistic regression; OOB: out-of-bag; RF: random forest.

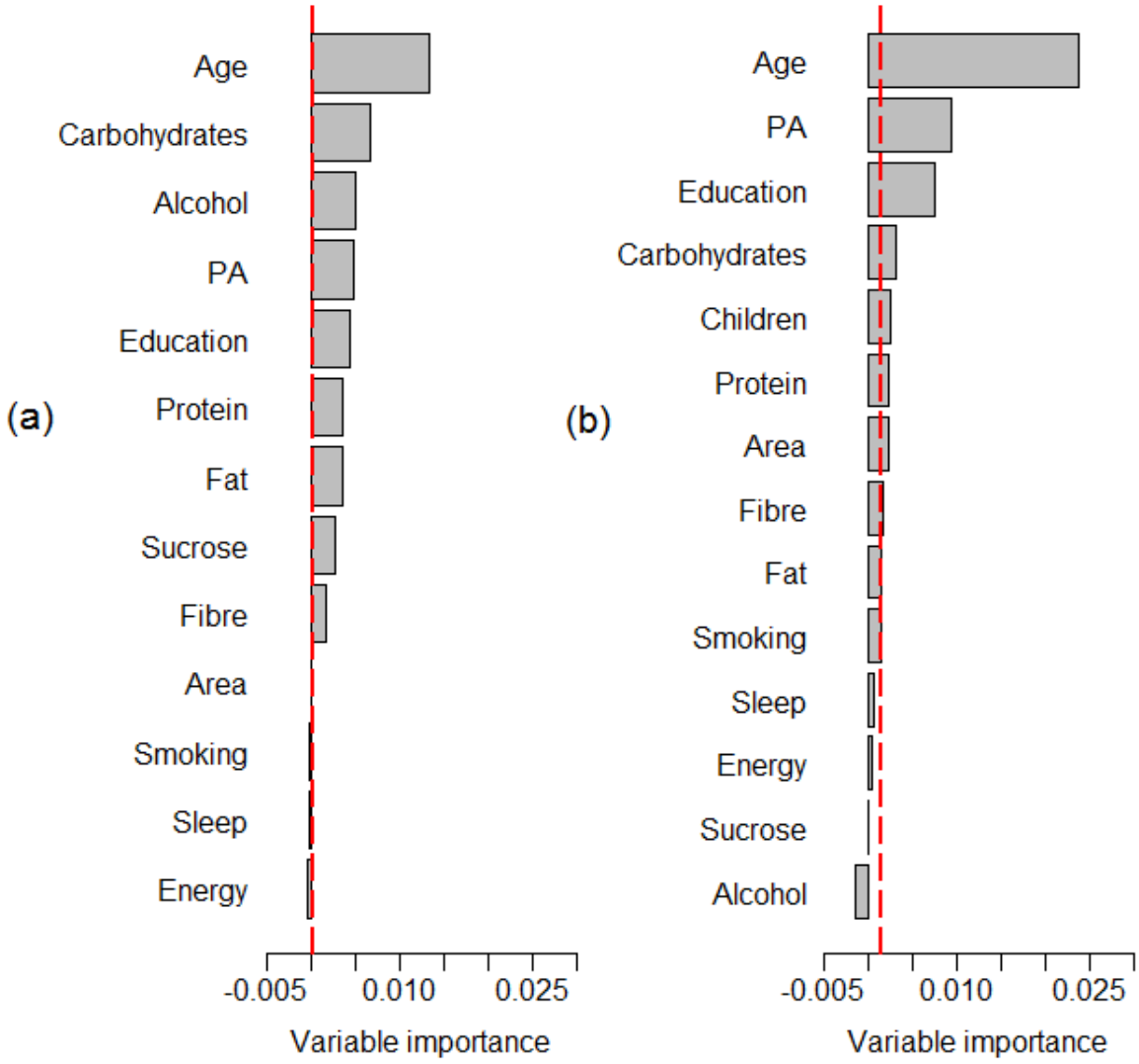
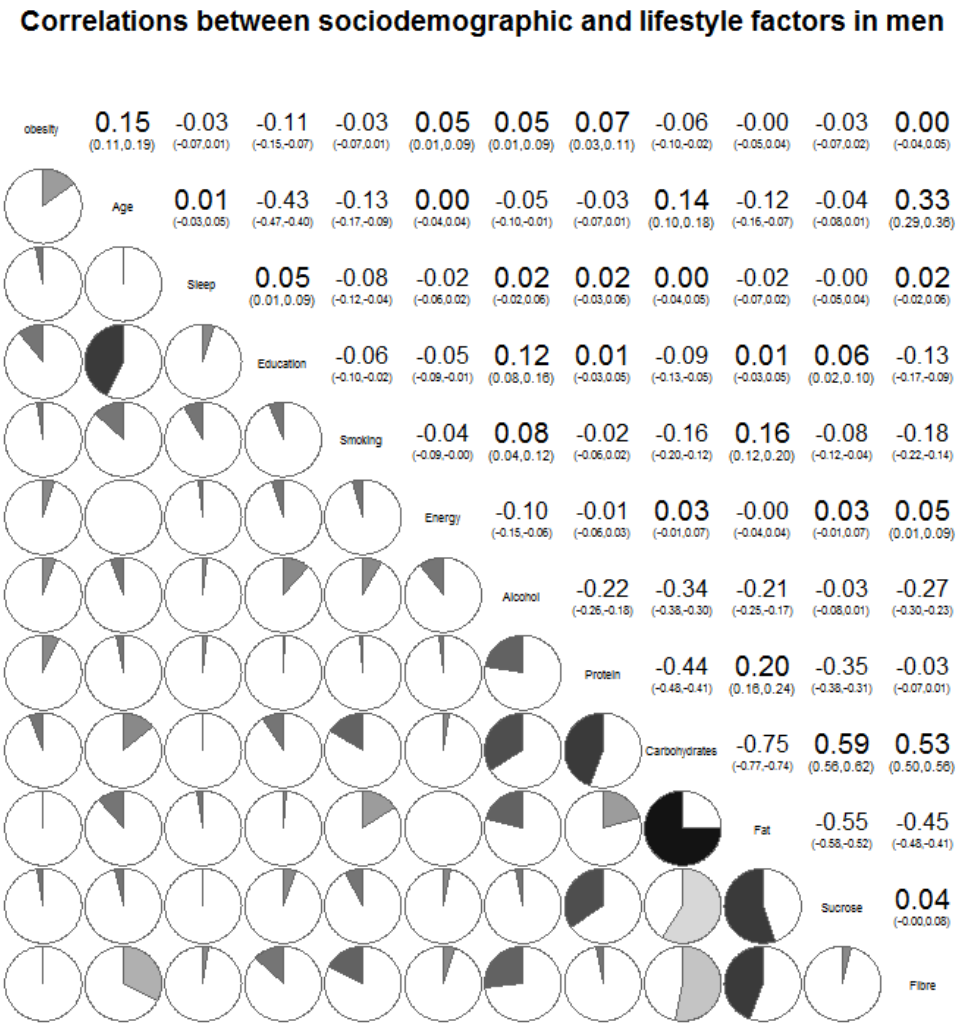
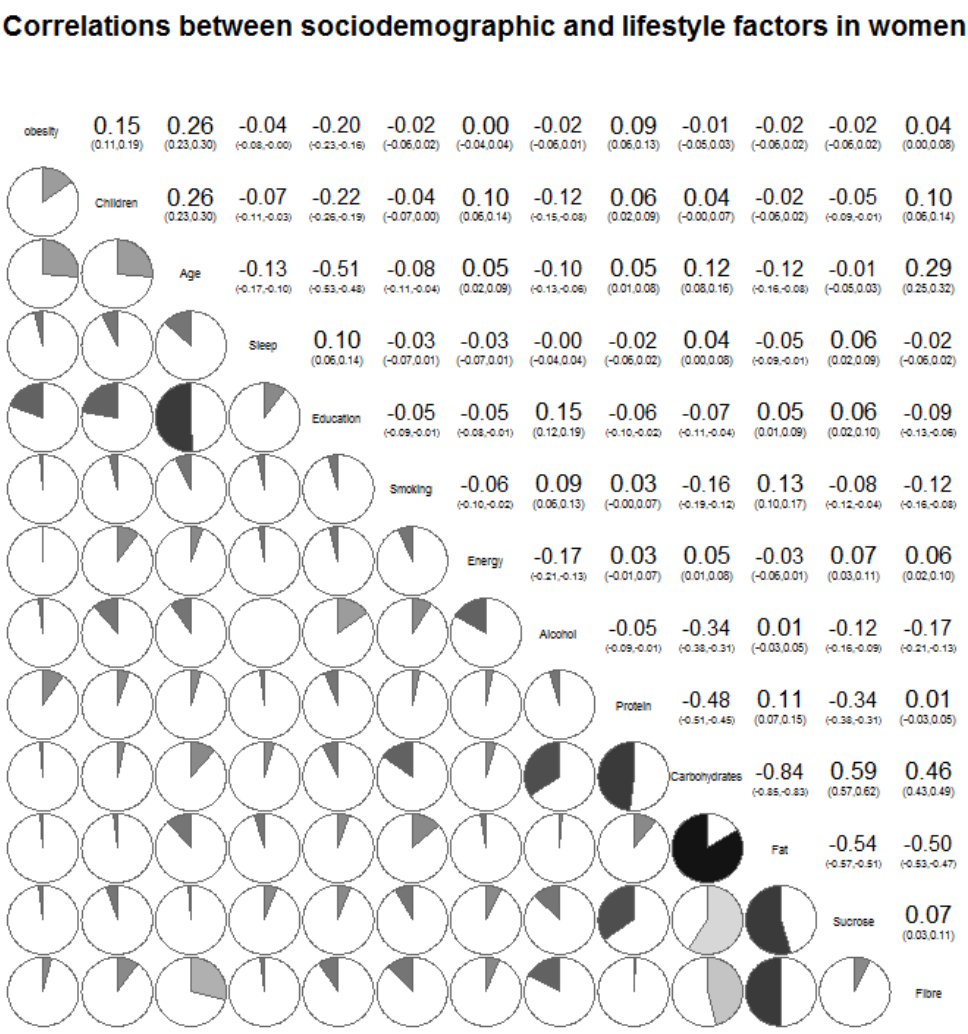


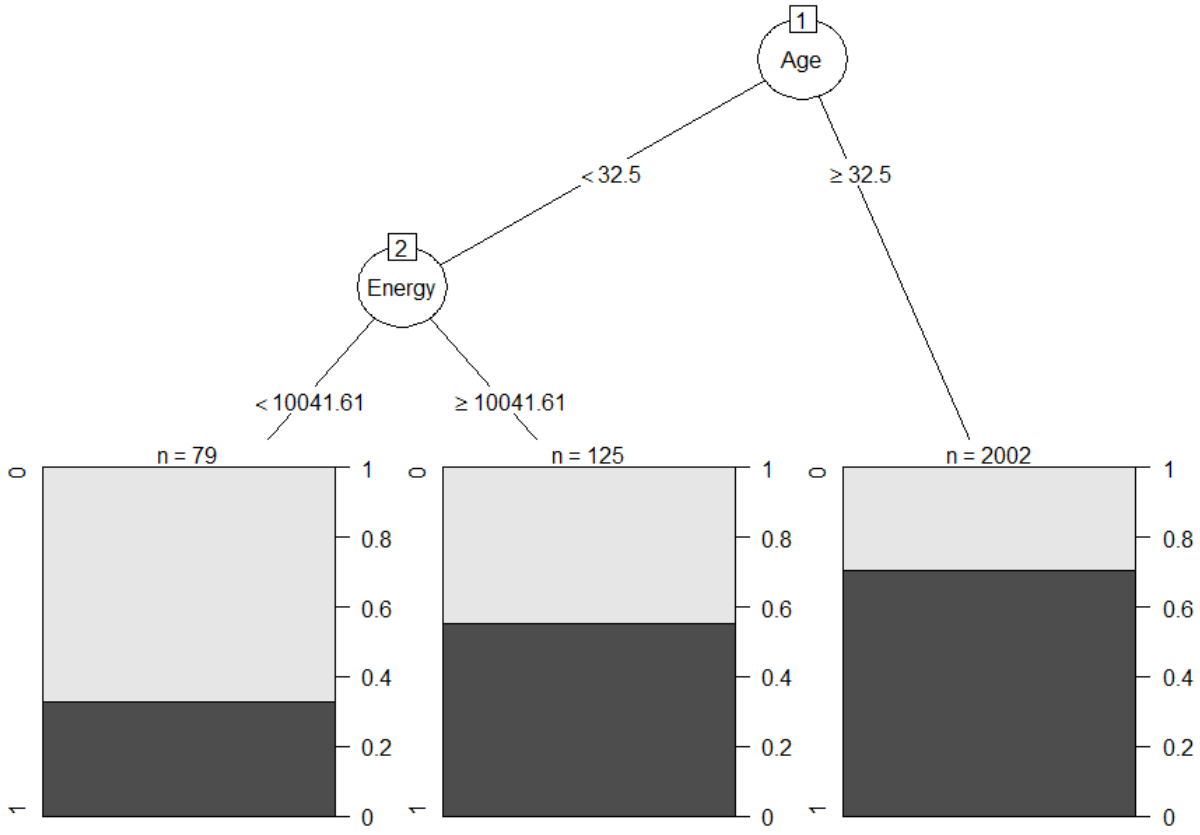
Figure 1. Permutation variable importance measures for men (*n* = 2206) (a) and women (*n* = 2551) (b) who participated in the National FINRISK 2007 Study. The larger the value the more important is the variable in reducing classification error. Variables which importance measure does not cross the dashed line are considered non-informative.



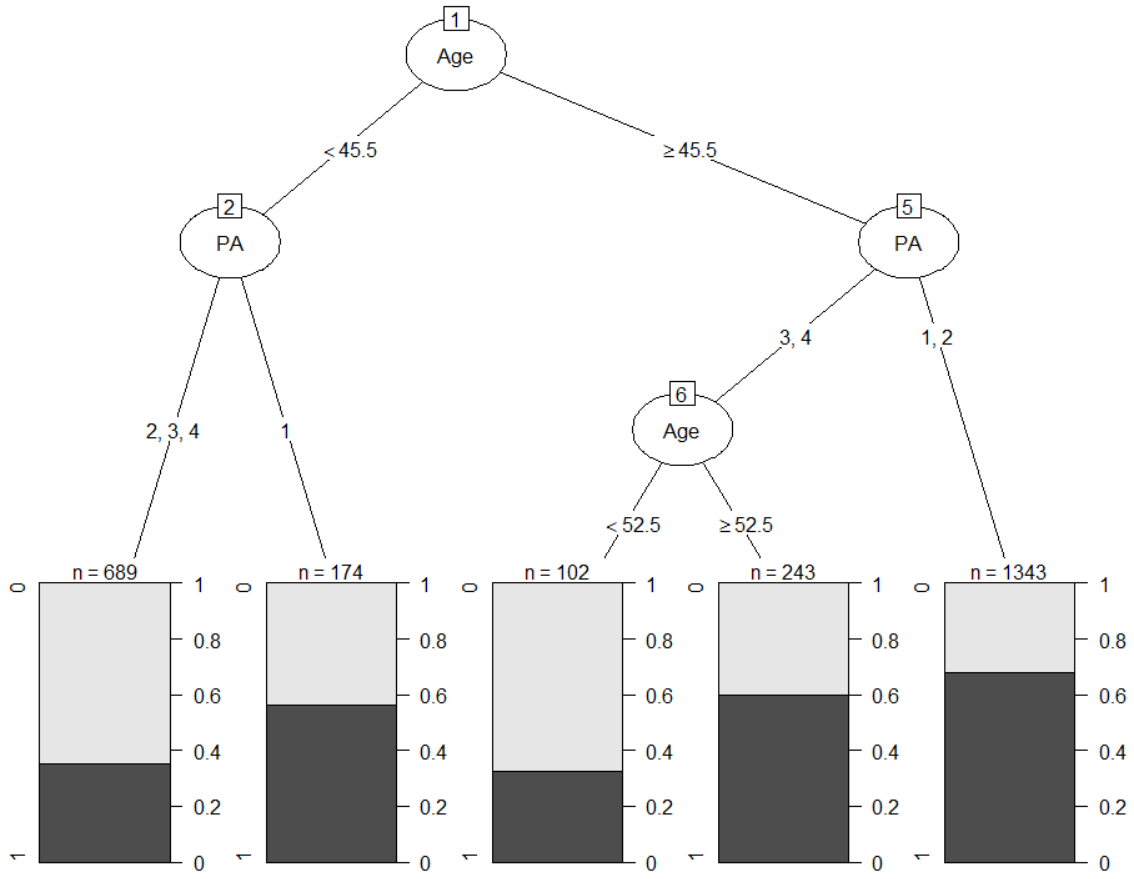
Supplemental Figure 1. Correlations between sociodemographic and lifestyle factors in men (*n* = 2206). Spearman's correlation coefficients and 95% confidence intervals are given in the upper triangle of cells (the cells above the principal diagonal). The lower triangle of cells illustrates the same information using pies. Here, the strength of the correlation is displayed by the size of the filled pie slice. The darker and more saturated the colour in the pie, the greater the magnitude of the correlation. Positive correlations fill the pie starting at 12 o'clock and moving in a clockwise direction. Negative correlations fill the pie by moving in a counterclockwise direction.



Supplemental Figure 2. Correlations between sociodemographic and lifestyle factors in women (*n* = 2551). Spearman's correlation coefficients and 95% confidence intervals are given in the upper triangle of cells (the cells above the principal diagonal). The lower triangle of cells illustrates the same information using pies. Here, the strength of the correlation is displayed by the size of the filled pie slice. The darker and more saturated the colour in the pie, the greater the magnitude of the correlation. Positive correlations fill the pie starting at 12 o'clock and moving in a clockwise direction. Negative correlations fill the pie by moving in a counterclockwise direction.



Supplemental Figure 3. Classification tree predicting overweight and obesity in men who participated in the National FINRISK 2007 Study (*n* = 2206). In each branch of the tree, nodes resulting with more normal-weight participants are classified to the left, and nodes resulting more overweight and obese participants are classified to the right. Predictor variables included in the model were age, living area, education, smoking status, leisure-time physical activity (PA), sleep duration, and intake of energy and macronutrients.



Supplemental Figure 4. Classification tree predicting overweight and obesity in women who participated the National FINRISK 2007 Study (*n* = 2551). In each branch of the tree, nodes resulting with more normal-weight participants are classified to the left, and nodes resulting more overweight and obese participants are classified to the right. Predictor variables included in the model were age, number of labours, living area, education, smoking status, leisure-time physical activity (PA), sleep duration, and intake of energy and macronutrients.