Genes Nutr (2008) 3:173–176
DOI 10.1007/s12263-008-0097-y

RESEARCH PAPER

Prediction of the metabolic syndrome status based on dietary and genetic parameters, using Random Forest

Fabien Szabo de Edelenyi · Louisa Goumidi · Sandrine Bertrais · Catherine Phillips · Ross MacManus · Helen Roche · Richard Planells · Denis Lairon

Received: 8 October 2008 / Accepted: 12 November 2008 / Published online: 26 November 2008

Abstract Metabolic syndrome (MS) is a cluster of metabolic abnormalities associated with an increased risk of developing cardio-vascular diseases, stroke or type II diabetes. Overall, the aetiology of MS is complex and is determined by the interplay between genetic and environmental factors although it is still difficult to untangle their respective roles. The aim of this study was to determine which factors and/or combination of factors could be predictive of MS status. Using a large case–control study nested in a well-characterized cohort, we investigated genetic and dietary factors collected at entry in subjects having developed MS 7 years later. We used a classification technique called Random Forest to predict the MS status from the analysis of these data. We obtained an overall out-of-bag estimation of the correct classification rate of 71.7% (72.1% for the control subjects and 70.7% for the cases). The plasma concentration of 16.1 was the most discriminative variable, followed by plasma concentration of C18.3(n-6) and C18.2. Three SNPs were selected by Random Forest (APOB rs512535, LTA rs915654 and ACACB rs4766587). These SNPs were also significantly associated to the MS by a univariate Fisher test.

Keywords Metabolic syndrome · Multivariate data analysis · Nutrigenomics · Random Forest

Introduction

Metabolic syndrome (MS) is characterized by the simultaneous presence of at least three of the following disorders: obesity (increased waist circumference), increased serum triglyceride level, increased fasting blood glucose, decreased HDL cholesterol level and high blood pressure. MS is significantly associated to elevated risk for cardio-vascular and metabolic diseases. The prevalence of the MS has dramatically increased over the last years, following the increase in the number of cases of obesity and the changes in the dietary habits of a larger part of the population [5]. The syndrome is probably due to the interplay between a genetic background and environmental factors such as dietary habits and lack of physical activity [3, 5]. However, it is still difficult to estimate the respective role of these factors in the onset of the MS.

In this study we wanted to determine which factors and/or combination of factors could be predictive of MS status. This could lead to a better understanding of the early stages of the MS and will also be useful preventing the MS.

F. Szabo de Edelenyi · L. Goumidi · R. Planells ·
D. Lairon (✉)
Faculté de Médecine Timone, UMR INSERM 476/INRA 1260,
27 bd Jean Moulin, 13385 Marseille, France
e-mail: Denis.Lairon@univmed.fr

S. Bertrais
UMR INSERM 557/INRA 1125/CNAM, Paris, France

C. Phillips · H. Roche
University College Dublin Conway Institute of Biomolecular
and Biomedical Research, Belfield, Dublin 4, Ireland

R. MacManus
Institute of Molecular Medicine, Trinity College Dublin, Dublin,
Ireland

Present Address:
L. Goumidi
Institut Pasteur de Lille, Unité d'Epidémiologie et de Santé
Publique, Inserm-UMR744, BP 245, 59019 Lille Cedex, France

Materials and methods

All our data were acquired on a large case–control study nested in a pre-existing cohort, the French SuViMax cohort. SuViMax is a randomized, double-blind, placebo-control, primary-prevention trial aimed to test the effect of anti-oxidant supplementation on the incidence of cancer and coronary heart diseases [4]. Due to the age of the included subjects, and since relevant parameters have been recorded during the 8 years follow-up of this cohort, this trial presented a particular interest for our study. Indeed during this time period, some subjects developed the MS, which thus provided us with a retrospective study of the development of the MS. Therefore, under these conditions, the study of the genetic susceptibility to the development of the metabolic syndrome can be achieved. Furthermore, reliable recording of dietary intakes and measurements of relevant markers made the study of diet-genes interactions possible.

The SuViMax-Lipgene cohort includes a total of 1,754 individuals (877 cases with MS and 877 matched controls). Case and control subjects have been matched according to age (58.2 ± 0.18) and gender. The scoring criteria used to select cases and controls are described in Fig. 1.

All individuals have been genotyped for 181 candidate genes resulting in 806 “tag” SNPs. The list of genes has been set up following an extensive literature study to find the genes potentially involved in MS. This involved searching through public databases including NCBIs, Pub Med and OMIM databank to draw up the list of candidate genes already associated with lipid metabolism, glucose homeostasis, insulin signalling and inflammation. Genes were selected from both functional physiological and population studies.

The choice of the SNPs for this list of genes has been performed using two different approaches. The first approach consisted to select SNPs that had already been studied in case–controls population or clinical studies. The second way to select SNPs of interest was to use bioinformatics techniques.

After DNA extraction from frozen stored samples, the genotyping has been performed by Illumina and KBioScience.

In addition to genetic data, we collected nutritional information and plasma fatty acid composition data at baseline. Nutritional data focused on the daily lipid intake. These data were acquired using several questionnaires. Six or more dietary inquiries were necessary to determine the average daily intake of different fatty acids.

After lipid extraction from stored plasma, methylation and gas chromatography, the plasma fatty acid composition was measured for each individual.

A classification technique called Random Forest was utilized to analyse our dataset and also for variable selection and MS status prediction. This technique has been proposed originally by Breiman and Cutler [1] and considers an ensemble of decision trees. A decision tree is a rule-based classifier using a succession of rules to iteratively split the data in subgroup. At each node, the most discriminative variable is found as well as a cut-off value and the dataset is split in two parts. Decision trees possess some advantages compared to many statistically based

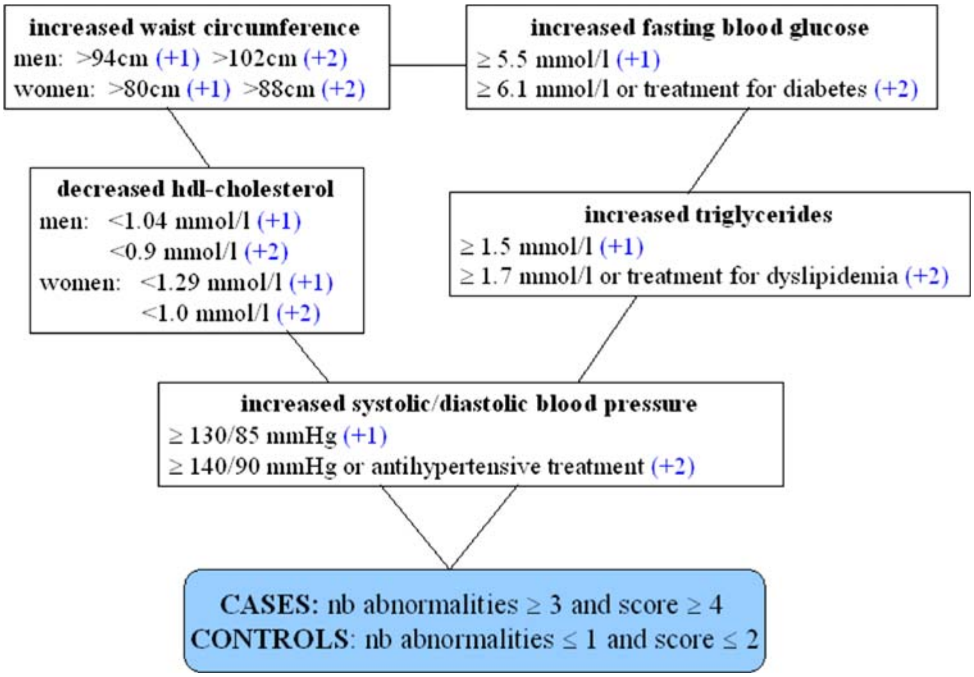
increased waist circumference
men: >94cm (+1) >102cm (+2)
women: >80cm (+1) >88cm (+2)
increased fasting blood glucose
≥ 5.5 mmol/l (+1)
≥ 6.1 mmol/l or treatment for diabetes (+2)
decreased hdl-cholesterol
men: <1.04 mmol/l (+1)
<0.9 mmol/l (+2)
women: <1.29 mmol/l (+1)
<1.0 mmol/l (+2)
increased triglycerides
≥ 1.5 mmol/l (+1)
≥ 1.7 mmol/l or treatment for dyslipidemia (+2)
increased systolic/diastolic blood pressure
≥ 130/85 mmHg (+1)
≥ 140/90 mmHg or antihypertensive treatment (+2)
CASES: nb abnormalities ≥ 3 and score ≥ 4
CONTROLS: nb abnormalities ≤ 1 and score ≤ 2

Fig. 1 Scoring criteria to select cases and controls

classification algorithms. First, they can easily manage dataset with a mixture of categorical variables (such as genotype information) and quantitative variables. They can also deal with a very large number of input variables, even when the number of variables is higher than the number of individual in the dataset. However, individual decision trees are not robust and the result of the classification depends highly of the dataset. The Random Forest technique has been developed to overcome this lack of robustness. The principle is to train a set of decision trees (typically several hundred trees) using different bootstrap samples from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the individual tree (sampling with replacement). Another characteristic of Random Forest is that only a subset of the input variable is used at each node to split the data. If there are M input variables, a number $m \ll M$ is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is constant and has to be defined as a parameter. We tested different values of m in order to optimize the classifier results and we found that the default value equal to sqrt(M) was given the best results.

To define the output of the Random Forest, the classification results from each tree are compiled using a majority of votes rule.

In Random Forest, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally using an out-of-bag estimation obtained by the classification of the individual left-out. The out-of-bag estimation has proven to be unbiased in many tests. It was then unnecessary to split our dataset into a training set and a test set.

The presence of missing values turned out to be a troublesome matter. Even though individual variables did not usually exhibit a large number of missing values (typically less than 10%), the combination of missing information over more than 800 variables decreased dramatically the number of complete individual. In order to maximize the number of individual available in the dataset, missing values have been imputed using two different algorithms. Missing genetic information has been estimated using an in-house program developed by David Tregouet (unpublished) which is using haplotype information and an EM algorithm to recover missing genotype. For the non-genetic part of the data, missing values were estimated using a Random Forest based algorithm (function rfImpute from the R package Random Forest).

Even though Random Forest can deal with a large number of input variables, a variable selection was necessary prior to the classification process in order to reduce the noise coming from uninformative variables. It allows determining the most informative variables. For this purpose we used the function cforest from the R package party. A major advantage of this function is that it can produce an unbiased measure of variable importance even in the case of a mixture of categorical variables and continuous variables [6].

After the final subset of input variables has been defined, the Random Forest algorithm was trained on our dataset. We used a Random Forest with 5,000 trees. It could then be used to classify any individual of the dataset or any new individual as case or control. We obtained also an out-of-bag estimation of the error rate.

Statement of informed consent

All subjects gave their informed written consent to the study, which was approved by the ad hoc ethical committees (i.e., The Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale and the Commission Nationale de l'Informatique et des Libertés).

Results

The variable selection process led us to only a small subset of the original variables. A large majority of the 806 SNPs were left-out during the process. For the final classification process we decided to keep only three SNPs as the inclusion of more SNPs did not improve the error rate and since we wanted to keep the subset of variables as small as possible. Besides these 3 SNPs, 2 dietary fat variables and 11 plasma fatty acid data were kept in the analysis as well as the level of physical activity (Fig. 2). We found that the plasma concentration of palmitoleic acid was the most discriminative variable, followed by plasma concentration of GLA and linoleic acid.

Palmitoleic acid represents only a small fraction of the total plasma fatty acids (around 2%) but it is very powerful to discriminate between cases and controls. The plasma concentration in palmitoleic acid is significantly higher for individual with the MS compared to controls.

The 3 SNPs selected by Random Forest (APOB rs512535, LTA rs915654 and ACACB rs4766587) were also found significant in the preliminary univariate association study performed on the 806 SNPs. It can be noticed that these three SNPs are involved in the three different aspects of the MS: glucose homeostasis (APOB), lipid metabolism (ACACB) and inflammatory process (LTA).

APOB rs512535 A > G and ACACB rs4766587 A > G are associated with a higher risk to develop the MS while LTA rs915654 has a protective effect.

We obtained an overall out-of-bag estimation of the correct classification rate of 71.4% (Table 1). Error rates

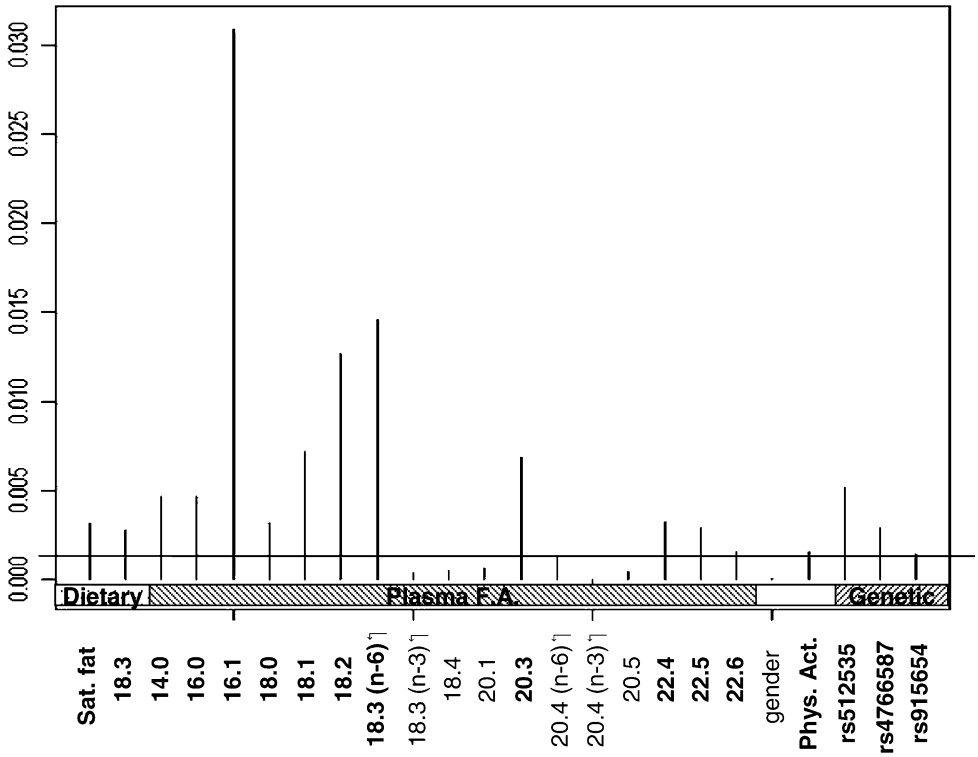

Fig. 2 Variable importance obtained by Random Forest for a part of the initial input variables. The 17 variables used for the classification process are displayed in *bold*

Table 1 Out-of-bag correct classification rates

	Predicted classes		Correct classification rates
	Control	Case	
MS status			
Control	632	245	72.1%
Case	257	620	70.7%
			Overall: 71.4%

for both groups were similar even though slightly lower for the control group. Cohen's Kappa value was calculated to estimate the agreement between the status predicted by Random Forest and the actual status [2]. We found a value of 0.428 which correspond to a moderate agreement (value between 0.41 and 0.6). However, this result is encouraging taken into account that the status was predicted 7 years before the actual onset of the syndrome and that no variables directly related to the MS such as BMI at entry were used for the classification. It can be noticed that when the BMI at entry is added as input variable in Random Forest, the error rate decreases dramatically (around 17%).

Conclusion

We developed a promising technique to predict the appearance of the MS several years in advance using some multivariate data analysis applied on dietary, environmental and genetic data. We found that some imbalance in the plasma fatty acid composition might reveal a risk of developing the MS.

Some genetic factors have been found to be significantly associated to the MS. However, the impact of the genetic background seems relatively limited compared to the influence of the dietary habits and environment.

Acknowledgments This work was supported by the European Commission, Framework Programme 6 (LIPGENE): contract number FOOD-CT-2003 505944.

References

1. Breiman L (2001) Random Forests. Mach Learn 45(1):5–32
2. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
3. Groop L (2000) Genetics of the metabolic syndrome. Br J Nutr 83:39–48
4. Hercberg S, Preziosi P, Galan P, Faure H, Arnaud J, Duport N, Malvy D, Roussel AM, Briançon S, Favier A (1999) "THE SU.VI.MAX STUDY": a primary prevention trial using nutritional doses of antioxidant vitamins and minerals in cardiovascular diseases and cancers. Food Chem Toxicol 37(9–10):925–930
5. Park YW, Zhu S, Palaniappan L, Heshka S, Carnethon MR, Heymsfield SB (2003) The metabolic syndrome: prevalence and associated risk factor findings in the US population from the Third National Health and Nutrition Examination Survey, 1988–1994. Arch Int Med 163:427–436
6. Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25