Non parametric bootstrap in factor analysis - white blood cell count and the metabolic syndrome.

Öhrvik, John
Karolinska Institutet, Department of Medicine
Cardiology Research Unit, N3:06
SE-171 76 Stockholm, Sweden
E-mail: john.ohrvik@ki.se

Introduction

Cardiovascular disease (CVD) is a major cause of morbidity and mortality in the developed world. As risk factors have been identified, more than one risk factor has been observed in many individuals. Such clustering of risk factors has been extensively studied; special interest has been focused on the clustering of risk factors in the Metabolic Syndrome (MetS). The MetS implies a clustering of certain cardiovascular risk factors in the same individual. According to most commonly used definitions this syndrome is defined by cut-off points in its principal variables: fasting glucose (FG), triglycerides (TG), high density lipoprotein cholesterol (HDL-c), waist circumference, and systolic/diastolic blood pressure (SBP/DBP).

The search for causative factors giving rise to the MetS has prompted the use of factor analysis to achieve an understanding of the fundamental pattern of clustering in the basic variables of the MetS. Factor analysis aims at ascertain whether the interrelation between a set of directly measurable variables are explicable in terms of a smaller number of underlying unobservable variables representing unique statistically uncorrelated domains termed *factors* of possible etiological importance. In middle aged people previous factor analyses of the continuous variables of the MetS have identified 2-4 essential factors.

Using factor analysis of variables of the MetS among 75-year-olds in a community based sample, we recently identified two factors of the MetS [1]. Factor 1, a metabolic factor, consisted of FG, TG, HDL-c and waist, whereas factor 2, a blood pressure factor, consisted of DBP and SBP. The clustering between the component variables of the metabolic factor was stronger among women than among men. Further, the metabolic factor was related to a decreased 10-years survival [1].

Markers of inflammatory processes and the MetS are related. Increased blood concentration of markers of inflammation, including white blood cell (WBC) count, is associated with high mortality and cardiovascular morbidity.

The purpose of the present study was to determine the relation between mortality and factors derived by a common factor analysis of WBC count and the basic components of the MetS. Nonparametric bootstrap was used to assess the consistency and accuracy of the factor analysis.

Study Population

The city of Västerås (130,000 inhabitants) situated in central Sweden has a population considered socioeconomically representative for the country. In 1997 a random sample of 618 of the 1,100 inhabitants born in 1922 (i.e. 75 years old) were invited to a cardiovascular health survey. The final number of participants were 432 (70% of those invited; women = 222; men = 210). Reasons for non-participation were unknown (n=46) or distributed on the invited as follows: never reached (n=29), died before examination (n=2), language or logistical problems (n=27), locomotive impairment (n=28) or unwilling due to diseases under treatment (n=54). Due to missing values the examined cohort finally comprised 196 men and 200 women (64% of the originally invited individuals). The study was approved by the research ethics committee at Uppsala University, Sweden.

Baseline examinations

Waist circumference was measured in the horizontal plane at the midpoint between the lowest rib and the iliac crest. Blood pressure (BP) was measured with a mercury sphygmomanometer and rounded to the nearest 5mmHg, with the subjects in a supine position and having relaxed for five minutes, which was the clinical practice in Sweden at that time. Diagnosis of previous myocardial infarction, stroke and diabetes was based on self-reported history of disease verified by medical records. Hypertension was defined as self-reported physician-diagnosed high blood pressure in combination with regular antihypertensive treatment.

The blood samples for FG, TG, HDL-c and WBC were collected in the morning with the subjects in a fasting state, details have been described elsewhere [1]. Newly detected diabetes was defined as FG ≥7.0 mmol/L without known diabetes.

The Metabolic Syndrome (MetS) was defined according to the National Cholesterol Education (NCEP) Adult Treatment Panel III (ATP III) criteria [2].

Follow up

Follow up started at the date of the index examination of the individuals in 1997. The participants were followed until death or to December 31, 2007. The follow-up was based on the Swedish Death Register. Assessment of follow-up survival status was complete in all participants.

Statistical methods

Continuous variables were summarized by median and quartiles and categorical variables by counts and proportions. For continuous variables the Wilcoxon Mann-Whitney rank sum test was used to compare groups. Categorical variables were compared using Fisher's exact test. The factor analysis included the WBC count and the individual variables of the MetS (FG, TG, HDL-c, Waist, DBP and SBP). Skewed variables (FG, TG and WBC count) were log-transformed prior to the factor analysis. To simplify the interpretation HDL-c was inverted (HDL-c^{-1}) prior to the factor analysis since high levels are protective as opposed to the other variables of the MetS.

Principle component analysis (PCA) within each sex strata was used to identify the initial set of uncorrelated factors. Since the variables were measured in different units the correlation matrix, which give each individual variable the same weight, was used in the analysis. As a result the sum of the eigenvalues, $\lambda_{i,}$ equals the number of variables. We applied the Kaiser-Guttman method, which implies that factors with $\lambda_i>1$ are retained since they summarize more information than any single variable. Unfortunately, a PCA of randomly generated uncorrelated variables will produce $\lambda_i>1$due to random fluctuation. To handle this sampling variation we applied non-parametric bootstrap resampling of individuals to assess the variability of the eigenvalues and construct confidence interval (CI). The number of bootstrap replicates *B* was set to 10,000. The percentile method, which uses the central (1-2α) part of the bootstrap distribution as the approximate (1-2α)100% CI was applied.

To facilitate the interpretation of the selected factors the varimax (orthogonal) rotation was used. This has as its rationale the provision of uncorrelated factors with a few large loadings and as many near-zero loadings as possible. Cut-off for factor loadings was set at 0.3. The factor loadings are the correlations between the original variables and the factors. Non-parametric bootstrap resampling of individuals was used to assess the consistency and accuracy of the factor analysis applying varimax rotation. A complication when applying bootstrap methods in factor analysis is that factor loadings are not uniquely defined because the sign of the loadings is arbitrary (i.e. the loadings for a factor can be multiplied by -1 without affecting the model estimates). The solution is to optimally reflect the resampled loadings towards e.g. the loadings of the original sample. Without reflection the resampled loadings of e.g. two factors will be randomly located in each of the four quadrants in the coordinate system spanned by the two factors. A further complication is that the order of the factors can change from one bootstrap sample to another, i.e. the factor connected to the largest eigenvalue will not be the same over all bootstrap replicates. The consequence is that the parameter estimates may not be smooth functions of the sample. A small perturbation of the sample can lead to a large change in the parameter estimates. To overcome this it has been suggested to apply Procrustes rotation [3]. We prefer to apply optimally reflected varimax rotation and use the number times the factor order differs from that in the original sample as a measure of the stability of the factor loadings. However when constructing confidence intervals based on e.g. the percentile method using the central (1-2α) part of the bootstrap distribution of the parameter estimate one has to assure that α is larger than the proportion of times the factor order differs from that in the original sample, otherwise large jumps can influence the length of the confidence interval.

Crude and adjusted prospective associations between all-cause mortality and the factors derived from the factor analysis were analyzed using Cox proportional hazards regression (PHREG) with and without stratification by sex. A best subset approach, using the corrected Akaike information criterion, $AIC_c = -2\log[L(\theta \mid \boldsymbol{x})] + 2\,(k+1)\,n\,/(n-k-2)$, where $L(\theta \mid \boldsymbol{x})$ is the likelihood function, k the number of estimated parameters and n the number of observations, as performance measure, was used to find an

'optimal' set of significant confounders among established cardiovascular risk factors with individual *p*-value < 0.20.

The predictive ability of the original component variables of the MetS, WBC count and the derived factors was assessed by AIC_c and the time dependent area under the receiver operating characteristic (ROC) curve, $AUC_t = P[Z_i > Z_j \mid D_i(t) = 1, D_j(t) = 0]$, where Z_i and Z_j are independent predicted risk scores, under the Cox PHREG model, for subject *i* and *j* and $D.(t) = 0/1$ is an indicator variable indicating whether an event has occurred by a specific time t [4].

The proportional hazard assumption was assessed by visual inspection of the log(-log(cumulative survival)) for each variable, continuous variables categorized into tertiles. Cumulative survival was estimated by the Kaplan-Meier estimator.

Results and Discussion

Table 1 presents baseline characteristics of the study population stratified by gender. Notably, MetS was significantly more common among women. Bootstrapped principal component analysis including the WBC count and the individual variables of the MetS identified three factors in men and two in women applying the Kaiser-Guttman method, see Figure 1. The three factors in men explained (95%CI within brackets) 66% (63-70%) of the total variation (1st factor 28% (25-32%), 2nd factor 23% 9 (20-25%) and 3rd factor 15% (14-17%)) and in women 57% (53-61%) of the total variation (1st factor 34% (30-37%) and 2nd factor 23% (21-25%)). After varimax rotation and optimal reflection (see Figure 2 for bootstrap clouds of the factor loadings) the 1st and 2nd factor in women were interpreted as a metabolic factor with significant loadings for log(FG), log(TG), HDL-c^{-1}, Waist and log(WBC) and near zero loading for DBP and SBP and as a blood pressure factor with highly significant loadings for DBP and SBP and near zero loadings for the other variables, see Table 2. This result was stable over all 10,000 bootstrap replicates. In men the results were not as stable which can be seen both from the Box plots of the eigenvalues in Figure 1 and the bootstrap clouds of the factor loadings in Figure 2. Also in men the 1st factor was the metabolic factor (in 9,443 of the 10,000 bootstrap replicates), but here without log(WBC), and the 2nd factor was the blood pressure factor (in 9,430 of the 10,000 bootstrap replicates), see Table 2. The 3rd factor in men can be interpreted as an inflammatory factor (in 9,464 of the 10,000 bootstrap replicates) with median loadings clearly above the cut-off for log(FG) and log(WBC), however the CI for loading of log(FG) covers zero. The individual factor scores were estimated using the factor loadings in bold in Table 2. In detail for each individual the standardized values of the original variables were multiplied by their corresponding factor loadings divided by the sum of square of the loadings for that factor.

During a median follow-up of 10.6 years (range 0.2-10.9), 145 individuals (37%) died (90 men 46% and 55 women 28%). The sex difference in mortality was highly significant ($p<0.001$); for men 5.4 deaths/100 person-year at risk and for women 2.8 deaths/100 person-year at risk. The main causes of death were cardiovascular (40 men; 27 women) and malignancy (27 men; 11 women). Ten year mortality among the185 invited individuals (89 men; 96 women) who did not participate in the study was considerably higher; 66 (74%) among men and 44 (46%) among women.

The metabolic factor was significantly related to 10-year mortality in both sexes and the inflammatory factor in men, see Table 3. The blood pressure factor had opposite non significant effects in men and women being negatively related with mortality in women. The interaction between the blood pressure factor and sex was nearly significant ($p=0.085$). In a pooled analysis the prognostic impact of the metabolic factor was only little attenuated upon adjustment for known hypertension, previous myocardial infarction and current smoking, the only significant confounders using AIC_c as performance measure in best subset approach, see Table 4. The predictive ability, measured as AIC_c and $AUC_{t=10\ yrs}$ were almost equivalent for the metabolic factor, HDL-c^{-1}, log(WBC) and the models including the individual variables of the metabolic factor with and without log(WBC), see Table 4.

Conclusions

The use of nonparametric bootstrap in factor analysis in combination with optimally reflected varimax rotation proved to be a useful method to assess the stability of the factor loadings. Further taking the instability into account standard bootstrap methods based on the bootstrap distribution of the parameter estimate could be used to construct confidence intervals. The closer relation between the individual components in women manifests itself in shorter confidence intervals for the factor loadings.

The key message of the present study is that in women, the WBC count is closely associated with

the metabolic variables FG, TG, HDL-c and waist, but not with the blood pressure variables. This association is not found in men where the WBC count and FG constituted a factor of their own. Further the association between the metabolic variables is much closer in women compared to men as can be seen from the bootstrap clouds in Figure 2 and the shorter confidence intervals in Table 2. The separation between the factors is also smaller in men leading to different order of the factors with respect to size of eigenvalue.

The present study shows that the metabolic factor significantly related to 10-year survival in both sexes, a relation which remained after adjusting for significant cardiovascular risk factors. The prognostic ability of the derived factors was not higher than that of the original variables and in particular HDL-c is a strong predictor of death. The clearly separated metabolic and blood pressure factors with disparity in prognostic impact may call in question the assumption of a unifying patho-physiology underlying MetS.

Strengths and limitations of the study

The restriction of our investigation to one age class enables us to leave out age as a confounding factor, creating the possibility of a meaningful analysis of the clustering of the MetS components and WBC count despite the limited number participants of the study. Furthermore, because of high participation rate, the participants are more representative for the population in a defined geographical area than described in most other studies on this topic.

These advantages are, however, obtained at the cost of the difficulties to generalize our findings to individuals not being 75-year old and to people from other geographical areas. However, it seems likely that our results are applicable to North European and white North American people in their seventies. A further limitation of the study is the fact that the mortality among invited individuals who did not participate in the study (36%) was considerably higher than among invited individuals who participated (64%), mainly reflecting a higher prevalence of diseases under treatment among non-participants.

Table 1 Baseline characteristics of the study population stratified by sex. Data are expressed as median (interquartile range) and number (%)*.

Variable	Men (n=196)	Women (n=200)	*p*-value
Fasting glucose, mmol/L	5.82 (5.40-6.49)	5.93 (5.48-6.49)	0.36
HDL cholesterol, mmol/L	1.36 (1.17-1.54)	1.62 (1.34-1.96)	na
Triglycerides, mmol/L	1.51 (1.11-1.92)	1.43 (1.11-2.07)	0.94
Waist, cm	94 (89-100)	88 (80-97)	na
Diastolic BP, mmHg	83 (80-90)	85 (80-90)	0.84
Systolic BP, mmHg	160 (150-180)	165 (150-190)	0.007
WBC count, 10^9/L	6.3 (5.4-7.2)	5.7 (4.8-6.8)	<0.001
Present MetS acc to NCEP III	48(24)	75(38)	0.007
High BP‡	118 (60)	128 (64)	0.47
Newly detected diabetes ‡‡	20 (10)	21 (11)	1.00
Current smoker	24 (12)	14 (7)	0.089
Cardiovascular disease	49(25)	30(15)	0.017
Previous myocardial infarction	30 (15)	9 (5)	<0.001
Angina pectoris	32 (16)	20 (10)	0.075
Stroke/TIA †	3 (2)	7 (4)	0.34
Heart failure	14(7)	12(6)	0.84
Known hypertension	52 (27)	58 (29)	0.65
Known diabetes	15 (8)	14 (7)	0.85

* HDL denotes high-density lipoprotein, BP blood pressure, WBC white blood cell, MetS Metabolic Syndrome and NCEP National Cholesterol Education Program. To convert the values for glucose to mg/dL divide by 0.0555. To convert the values for cholesterol to mg/dL divide by 0.0259. To convert the values for triglycerides to mg/dL divide by 0.0113.

† TIA, transient ischemic attack.

‡High BP is defined as SBP≥140 mmHg or DBP≥90 mmHg without known hypertension.

‡‡ Newly detected diabetes is defined as fasting glucose ≥7.0 mmol/L without known diabetes.

na – not applicable.

Table 2 Median factor loadings with bootstrap percentile intervals (89% for men, since the factor order is changed in around 5.5% of the replicates and 95% for women) based on 10,000 replicates. Loadings of the individual components included in the respective factor in bold (cut-off = 0.30). Only for the inflammatory factor the sample loadings differed from the median loadings with more than 2 units in the 2nd decimal 0.49 vs 0.44 and 90 vs 0.87.

	Median factor loadings with bootstrap percentile interval (89% for men and 95% for women)				
Component	Metabolic factor		Blood pressure factor		Inflammatory factor
	Men	Women	Men	Women	Men
log (FG)	**0.35** (0.04-0.60)	**0.67** (0.57-0.74)	-0.03 (-0.23-0.26)	-0.06 (-0.24-0.12)	**0.44** (-0.59-0.74)
HDL-c^{-1}	**0.78** (0.56-0.84)	**0.80** (0.73-0.85)	-0.10 (-0.23-0.15)	-0.03 (-0.19-0.14)	-0.01 (-0.20-0.17)
log (TG)	**0.83** (0.66-0.87)	**0.75** (0.65-0.81)	-0.05 (-0.17-0.20)	0.00 (-0.18-0.19)	0.15 (-0.05-0.36)
Waist	**0.61** (0.36-0.72)	**0.67** (0.53-0.76)	0.26 (0.09-0.50)	0.16 (-0.05-0.40)	-0.08 (-0.45-0.27)
Diastolic BP	0.06 (-0.05-0.25)	0.04 (-0.06-0.13)	**0.87** (0.70-0.90)	**0.89** (0.85-0.92)	-0.10 (-0.22-0.02)
Systolic BP	-0.07 (-0.17-0.12)	-0.02 (-0.12-0.07)	**0.85** (059-0.89)	**0.87** (0.82-0.91)	0.15 (0.03-0.28)
log (WBC)	0.00 (-0.14-0.22)	**0.50** (0.29-0.64)	0.07 (-0.06-0.25)	0.00 (-0.25-0.29)	**0.87** (0.66-0.97)

Table 3 Hazard Ratios (HR) and 95% CIs for all-cause mortality per 1 unit increase.

Model	AIC_c	*p*-value	HR (95%CI)
Metabolic factor men		0.007	1.22 (1.06-1.41)
Metabolic factor women		0.010	1.25 (1.06-1.48)
Blood pressure factor men		0.20	1.12 (0.94-1.33)
Blood pressure factor women		0.25	0.88 (0.71-1.09)
Inflammatory factor men		0.009	1.29 (1.07-1.56)
Metabolic factor adjusted for sex	1541.9	<0.001	1.23 (1.11-1.38)
Blood pressure factor adjusted for sex (men=0, women=1) Interaction sex*Blood pressure factor	1554.4	0.18 0.085	1.13 (0.95-1.34) 0.79 (0.60-1.03)

Table 4 Adjusted Hazard Ratios (HR) and 95% CIs for all-cause mortality per 1 unit increase.

Model	AIC_c	$AUC_{t=10\ yrs}$	*p*-value	HR (95%CI)
Metabolic factor†	1506.1	0.697	0.010	1.16 (1.04-1.29)
HDL-c^{-1} †	1503.8	0.700	0.003	4.25 (1.65-10.95)
log(WBC) †	1505.8	0.690	0.010	2.70 (1.27-5.71)
log(FG), HDL-c^{-1}, log(TG), Waist†	1506.2	0.707	0.014‡	
log(FG), HDL-c^{-1}, log(TG), Waist, log(WBC)†	1503.9	0.713	0.008‡	
Blood pressure factor (men=0, women=1) Interaction sex*Blood pressure factor†	1511.6	0.676	0.32 0.086	1.10 (0.91-1.32) 0.78 (0.59-1.04)
Inflammatory factor men††	-	0.678	0.055	1.20 (1.00-1.44)
Adjusting variables	1510.5	0.583	<0.001‡‡	

†Adjusted for sex, known hypertension, previous myocardial infarction and current smoking.
††Adjusted for known hypertension, previous myocardial infarction and current smoking.
‡*p*-value for the difference in Wald χ^2 between the full model and the model with only adjusting variables.
‡‡ *p*-value for the difference in Wald χ^2 between the model with adjusting variables and the null model.

REFERENCES

1. Ohrvik J, Hedberg P, Jonason T, Lonnberg I and Nilsson G. Factor Analysis of the Individual Components of the Metabolic Syndrome among Elderly Identifies Two Factors with Different Survival Patterns-A Population-Based Study. *Metab Syndr Relat Disord.* Mar 14 2009; 7(3):171-178.

2. NCEP. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *Jama.* May 16 2001;285(19):2486-2497.

3 Timmerman M. E., Kiers H.A.L. and Smilde A. G. Estimating confidence intervals for principal components loadinds: A comparison between the bootstrap and asymptotic results. *Brit Jour Math Statist Psychol,* 2007; 12(3):359-379.

4. Chambless L. E., Cummiskey C. P. And Gang C. Several methods to assess improvement in risk prediction models: Extension to survival analysis. *Statist Med,* 2010; DOI:10.1002/sim.4026

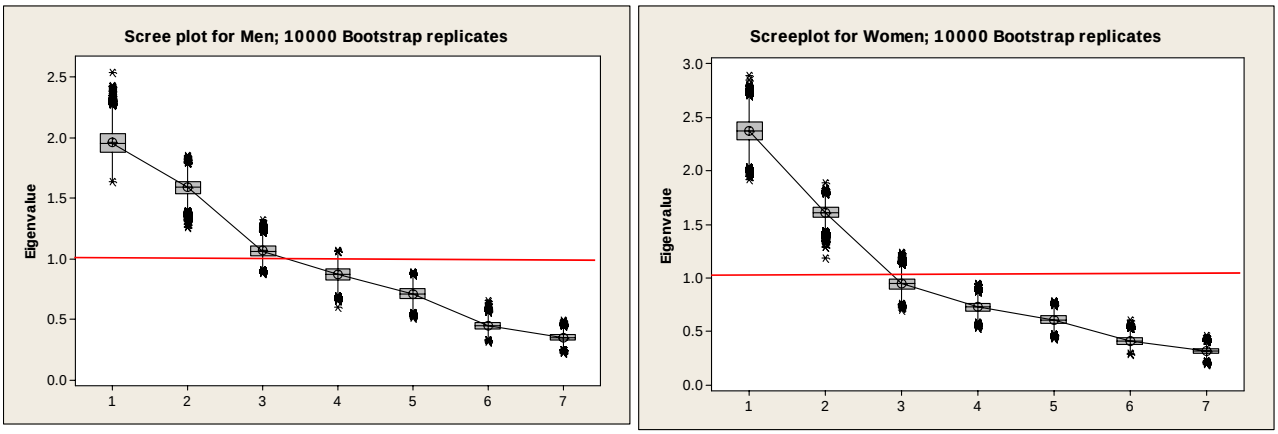

Figure 1 Scree plot presented as Box plots from 10,000 bootstrap samples. Men left; Women right.

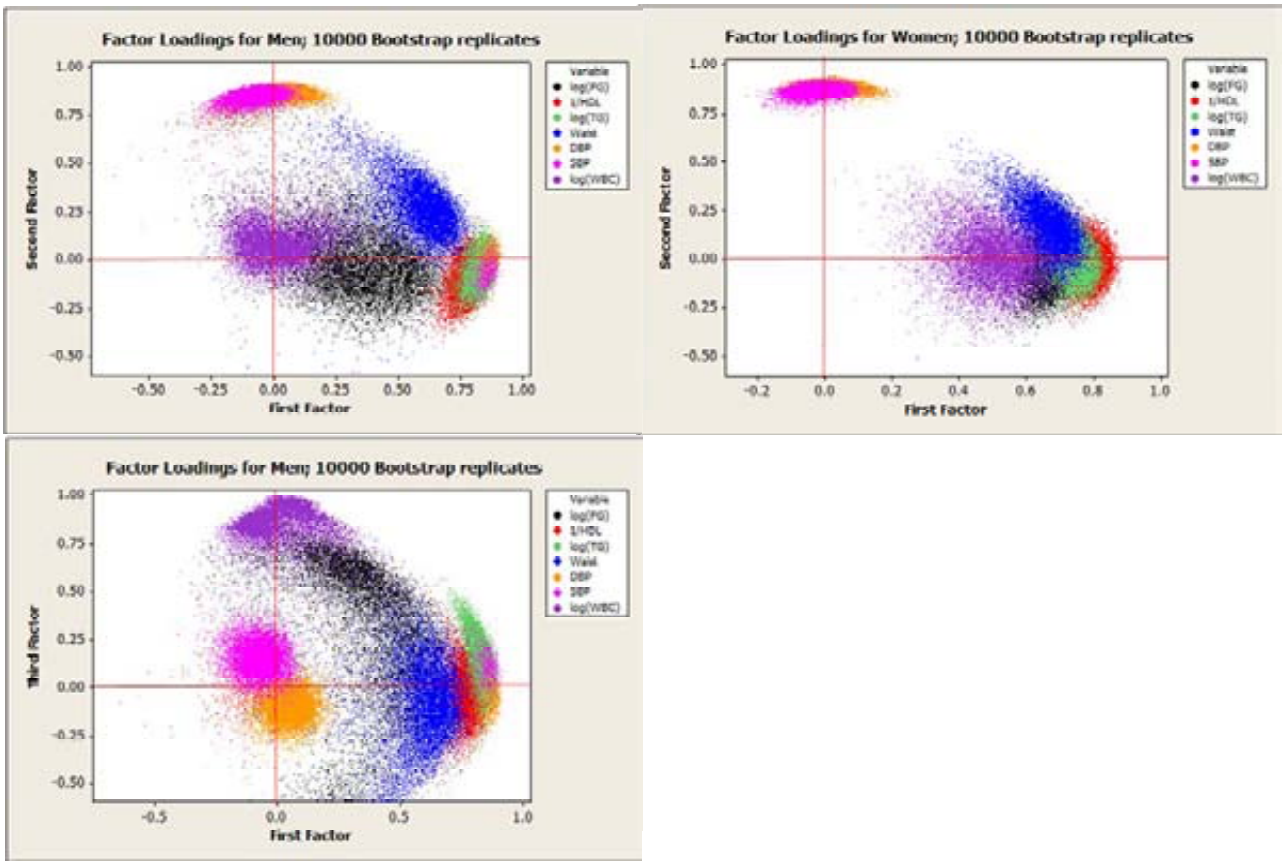

Figure 2 Bootstrap clouds (10,000 bootstrap samples) of the factor loadings for the individual components of the MetS in optimally reflected varimax rotated space. Men left; women right.