Published in final edited form as:
Epidemiology. 2016 November ; 27(6): 870–878. doi:10.1097/EDE.0000000000000548.

Classification and clustering methods for multiple environmental factors in gene-environment interaction – application to the Multi-Ethnic Study of Atherosclerosis

Yi-An Ko[a], **Bhramar Mukherjee**[b], **Jennifer A. Smith**[c], **Sharon L.R. Kardia**[c], **Matthew Allison**[d], and **Ana V. Diez Roux**[e]

[a]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

[b]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[c]Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109, USA

[d]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA 92093, USA

[e]Department of Epidemiology and Biostatistics, Dornsife School of Public Health at Drexel University, Philadelphia, PA 19104, USA

Abstract

There has been an increased interest in identifying gene-environment interaction (G×E) in the context of multiple environmental exposures. Most G×E studies analyze one exposure at a time, but we are exposed to multiple exposures in reality. Efficient analysis strategies for complex G×E with multiple environmental factors in a single model are still lacking. Using the data from the Multi-Ethnic Study of Atherosclerosis, we illustrate a two-step approach for modeling G×E with multiple environmental factors. First, we utilize common clustering and classification strategies (e.g., k-means, latent class analysis, classification and regression trees, Bayesian clustering using Dirichlet Process) to define subgroups corresponding to distinct environmental exposure profiles. Second, we illustrate the use of an additive main effects and multiplicative interaction model, instead of the conventional saturated interaction model using product terms of factors, to study G×E with the data-driven exposure sub-groups defined in the first step. We demonstrate useful analytic approaches to translate multiple environmental exposures into one summary class. These tools not only allow researchers to consider several environmental exposures in G×E analysis but also provide some insight into how genes modify the effect of a comprehensive exposure profile instead of examining effect modification for each exposure in isolation.

INTRODUCTION

There has been a growing interest in identifying gene-environment interaction (G×E) effects on quantitative traits associated with complex diseases in longitudinal cohort studies.[1] Most

Correspondence: Yi-An Ko, Department of Biostatistics and Bioinformatics, Emory University, 1518 Clifton Rd, Atlanta, GA 30322, USA, Phone: (404) 727-8128, yi-an.ko@emory.edu.

Conflict of interest: None declared.

G×E studies study one environmental exposure and its interaction with candidate genes.[2,3] In reality, we are exposed to multiple often intercorrelated factors, and they can jointly modify genetic effects. Efforts to study multiple exposures in isolation can result in underestimated environmental modifying effects. Identifying the modifying effect of a complete environmental exposure profile is important not only to understand the etiologic genetic role but to identify subgroups for targeted intervention to address clustered environmental factors.

Very few studies considered multiple exposures when investigating G×E,[4] possibly due to the challenges associated with variable selection, interpretation, and efficient statistical modeling. Existing analysis strategies for multiple environmental factors include an omnibus risk regression model (a full model that contains as many exposure variables as possible), a reduced model generated by some model selection methods (e.g., univariate thresholding, best subset, stepwise regression), and multilevel/hierarchical modeling methods.[5] Environmental epidemiologists have also used principal component analysis, factor analysis, or computing a risk score to reduce the dimensionality of the exposure measures.[6,7]

G×E studies are also challenged by multiple categories of exposures and genes. An interaction model including cross-product terms of gene (G) and environmental exposure (E) is typically used for testing G×E with quantitative traits.[8] When both G and E are treated as categorical variables (commonly used in epidemiologic practice),[9,10] a product form for G×E results in a saturated interaction structure. A saturated interaction structure estimates a parameter for each configuration of G and E without structural assumption for the interaction. The number of parameters and the degrees of freedom (df) for the interaction test grow with the number of G or E categories. This is not practical for finely cross-classified data and may yield inefficient parameter estimates and loss of power compared to a parsimonious model. Although several parsimonious interaction models for G×E have been proposed (e.g., Tukey's single df model for non-additivity[11–13]), they have limited power for detecting interaction if the model is misspecified.

The class of additive main effects and multiplicative interaction models,[14] frequently used in crop cultivar trials, provides a solution to the problem of modeling interaction in cross-classified tables with many categories. The additive main effects and multiplicative interaction model was proposed as the "FANOVA" (factor analysis of variance) model,[15,16] involving both additive and multiplicative components underlying a two-way data structure. The model first removes additive main effects and then applies principal component analysis to the residual to capture "non-additivity" under a new set of coordinate axes without imposing specific interaction structural assumptions. The interaction term can be approximated by a small number of leading multiplicative components, such that the effective df of the resultant G×E test is reduced. Addictive main effects and multiplicative interaction performs well across a spectrum of interaction structures.[12] It is useful for detecting interaction effects in the absence of main effects.[13]

We illustrate a two-step approach for G×E test with multiple environmental factors using data from the Multi-Ethnic Study of Atherosclerosis (MESA).[17] First, we describe four common clustering and classification strategies to synthesize information from multiple environmental factors and define exposure profile subgroups. We also introduce the additive main effects and multiplicative interaction model. Next, to illustrate our analytical framework using the MESA data, we consider body mass index (BMI) as our outcome. Exposures include dietary intake, physical activity, and psychosocial factors. We apply the clustering or classification techniques to these environmental factors and create an overall exposure profile for each study participant, which is essentially a single categorical variable with categories defined by the measured "environmental characteristics". Finally, we investigate G×E between BMI-related genes and the overall exposure profile using additive main effects and multiplicative interaction models and compare the results with saturated interaction and Tukey's one df models.

METHODS

Clustering and Classification Methods to Define Exposure Groups

K-means Clustering—K-means clustering is a nonparametric partitioning method seeking a minimum for the error sum of squares and is suitable for quantitative variables.[18,19] The k-means algorithm (1) randomly selects k centroids (k < number of data points) and assigns each data point to its closet centroid by minimizing the within-cluster sum of squares and (2) recalculates the centroids as the average of all data points in a cluster and again assigns data points to their closest centroids. Step (2) is continued until the observations are not reassigned or the maximum number of iterations is reached. The optimal number of clusters is chosen based on plotting the number of clusters and the corresponding total within-cluster sum of squares.

Latent Class Analysis—Latent class analysis detects latent classes so that observed variables can be explained by a single unobserved latent categorical variable based on their maximum likelihood class membership.[20] The latent classes divide individuals into mutually exclusive groups. For example, people are categorized based on their eating habits into different dietary patterns (latent classes). This could lead to finding diet categories, such as Western pattern diet, Mediterranean diet, etc. The R package **poLCA** performs traditional latent class analysis using categorical variables, while **mclust** performs latent class analysis using continuous data.[21] Model parameters are estimated using maximum likelihood, and the best normal mixture model is chosen according to Bayesian information criterion values among different covariance structures and different cluster numbers.

Classification and Regression Trees—Classification and regression trees are a supervised learning technique that recursively partitions data into smaller groups based on a categorical (for classification trees) or continuous (for regression trees) dependent variable and one or more independent variables (categorical or continuous) to enhance homogeneity within groups.[22] At each split, data are partitioned into two mutually exclusive groups based on an independent variable. The splitting procedure is applied to each group separately. Classification and regression tree generates a sequence of sub-trees by growing a large tree and pruning it back.[22] It sequentially collapses nodes that result in the smallest change in purity then uses cross-validation to select the optimal decision tree. **rpart** and **tree** are the primary R packages and can handle both categorical and continuous variables.

Bayesian Dirichlet Process Clustering—Bayesian Dirichlet process clustering discovers subgroups by allowing the number of groups to vary and links cluster membership to the outcome via a regression model (supervised clustering). This method consists of two submodels.[23] An allocation model (i.e., a discrete mixture model) is constructed to assign an individual to a cluster, which incorporates a Dirichlet process prior on the mixing distribution.[24,25] Once individuals are assigned into groups, the cluster profiles are used as a categorical predictor in a disease submodel for the outcome (with or without covariates) simultaneously. Markov chain Monte Carlo methods are used to fit the model that allows the number of clusters to vary between iterations of the sampler. At each iteration, a score matrix is created to indicate whether two individuals belong to the same cluster. A probability matrix **S** is obtained at the end of iterations by averaging the score matrices to denote the pairwise probabilities that individuals are assigned to the same cluster. Finally, an optimal number of clusters can be found by minimizing the least-squared distance to **S** among all the partitions explored by Markov chain Monte Carlo or processing **S** through partitioning around medoids (an algorithm very similar to k-means). We illustrate the method using the R package **PReMiuM**.[26]

Remarks—There are several important distinctions among the four clustering algorithms. First, the exposure profile groups defined by classification and regression tree and Bayesian Dirichlet process clustering are conditional on the outcome variable (supervised learning), whereas those defined by k-means and latent class analysis are not, indicating that the formation of classification and regression tree/Bayesian Dirichlet process groups assumes some existing association between the exposures and the outcome. When these exposure clusters are used in G×E analysis, one may run the risk of over-fitting and underestimating sampling variance. Given that the attempt is to capture an individual's exposure characteristics, it may be more meaningful to derive exposure clusters regardless of the outcome. Using unsupervised clusters at the first step may be more principled and desirable with the added advantage of the clusters being transportable across different health outcomes. Second, both latent class analysis and Bayesian Dirichlet process clustering have a solid statistical framework, while k-means and classification and regression trees are based on purely nonparametric algorithms. Nevertheless, the advantage of algorithm-based approaches lies in computational efficiency and as such, they may be preferable for large datasets. Third, standardized variables are required for k-means and latent class analysis, whereas classification and regression trees are invariant under monotone transformations of variables and are immune to outliers. Fourth, latent class analysis, classification and regression trees, and Bayesian Dirichlet process clustering can accommodate different data types, while the standard k-means algorithm can only handle numeric variables. Classification and regression trees tend to perform better for discrete/categorical features and tends to select categorical variables with many unique values for splits over ordinal variables[27]; they also feature an easily interpretable representation and are preferred in biomedical applications. Lastly, the primary advantage of Bayesian Dirichlet process clustering over traditional approaches with a fixed number of clusters is that the clustering uncertainty can be evaluated by re-examining the Markov chain Monte Carlo output. Regarding missing data, latent class analysis makes a missing at random assumption, but the other approaches assume data are missing completely at random. Otherwise, multiple imputation of missing values can be conducted prior to applying the clustering methods.

Interaction Models for G×E

A G×E model for the outcome Y includes main effects of G and E and the G×E interaction as well as potential confounding factors. Let y_{ij} be the j-th observation for subject i. A linear mixed model including G×E as a fixed effect in longitudinal studies can be expressed as

$$y_{ij}=\beta_0+\beta_g G_i+\beta_e E_{ij}+\beta_{ge} G_i E_{ij}+\mathbf{x}_{ij}^t\boldsymbol{\beta}+\mathbf{u}_{ij}^t\mathbf{b}_i+\varepsilon_{ij},\ i=1,\ldots,N, j=1,\ldots,n_i,$$

where G_i is the genetic factor for subject i, and E_{ij} is the j-th repeated measure of the environmental exposure on subject i, $\mathbf{x}_{ij}$ and $\mathbf{u}_{ij}$ are fixed-effects of other covariates and random-effects design matrix corresponding to the j-th observation for subject i; β_g is the coefficient for genetic effect, β_e is the coefficient for environmental effect, β_{ge} is the coefficient for interaction effect, $\boldsymbol{\beta}$ is a vector of covariate coefficients, $\mathbf{b}_i \sim N_q(\mathbf{0}, \boldsymbol{\Phi})$ and $\boldsymbol{\Phi}$ is the covariance matrix of random effects $\mathbf{b}_i$, and ε_{ij}~N(0, σ^2 is the measurement error, n_i is the number of observations for subject i, and N is the total number of subjects. The covariance structure for $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i1}, \ldots \varepsilon_{in_i})^t$ can be defined such that it accounts for within-subject correlation. When G and E are categorical variables with R and C categories, G×E is often analyzed in the form of a R × C table. Due to the sum-to-zero constraints for parameter identifiability, we have (R − 1)(C − 1) interaction parameter estimates and also (R − 1)(C − 1) df for testing in a fully saturated interaction model. The number of parameters and the df for testing interaction increases significantly with increased R or C, which may contribute to decreased efficiency and loss of power to detect interactions.

Tukey's One Degree-of-Freedom Model—Let $\mathbf{T}_{R\times C}$ be the interaction matrix (i.e., residual matrix after main effects and covariate effects removed) with τ_{rc} as the (r,c)-th element ((r=1, … ,R,c=1, … ,C). A Tukey's one-df interaction[28] has the form:

$$\tau_{rc}=\theta\beta_r^G\beta_c^E,$$

where β_r^G and β_c^E the parameters for genetic main effects and exposure main effects corresponding to the r-th and c-th categories of gene and exposure (with constraints $\sum_{r=1}^R\beta_r^G=\sum_{c=1}^C\beta_c^E=0$), and θ is a scale parameter for the interaction effect. Testing interaction is equivalent to testing $H_0{:}\theta = 0$. Since the interaction term in Tukey's model is a scaled product of main effects, the existence of interaction is conditional on the presence of main effects.

Additive Main Effects and Multiplicative Interaction Model—In additive main effects and multiplicative interaction models, a singular value decomposition is performed for the interaction matrix $\mathbf{T}_{R\times C}=\mathbf{AD\Gamma^T}$, where $\mathbf{A}$ and $\boldsymbol{\Gamma}$ are R × s and C × s orthonormal matrices (i.e., $\mathbf{A}^T\mathbf{A} = \boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}$) and $\mathbf{D}$ is a s× s diagonal matrix with elements $d_1 \geq d_2 \geq \ldots \geq$

d_s (s = min(R − 1, C − 1)). Gollob[15] proposed to retain the first M (M ≤ s) components of this representation,

$$\tau_{rc}=\sum_{m=1}^{M} d_m \alpha_{rm} \gamma_{cm}+\phi_{rc}.$$

The first few leading terms are considered as the signal of G×E, whereas the higher-order terms are regarded as random noise. This model was later named as additive main effects and multiplicative interaction (AMMI) models because the mean response was explained via additive main effects and multiplicative contrasts explaining the residual variation after fitting additive main effects.[29] When M = s, then $\phi_{rc} = 0$, and we have a fully saturated interaction model. When M < s, we have a lower rank representation of **T** and hence a reduced effective df for the interaction test. Given that three genotypes for each single nucleotide polymorphism (SNP) are considered (i.e., M ≤ 2), we focus on M = 1 in the additive main effects and multiplicative interaction model (calling it AMMI1), namely,

$$\tau_{rc}=d_1 \alpha_{r1} \gamma_{c1}+\phi_{rc}.$$

Concerning testing G×E of the above form, we adopt the parametric bootstrap method for additive main effects and multiplicative interaction to avoid computational iteration.[13] Briefly, a fully saturated interaction form is used for G×E modeling and then a singular value decomposition is applied to the estimated interaction matrix $\hat{\mathbf{T}}$. Then $\hat{d}_1$ can be approximated by the largest singular value, and $\hat{\alpha}_{r1}$ $\hat{\gamma}_{c1}$ are approximated by the corresponding left and right singular vectors. Subsequently, a pivot based on likelihood ratio test is constructed using the two-step regression estimates. The null distribution of this pivot is derived using parametric bootstrap.

G×E ANALYSIS

Multi-Ethnic Study of Atherosclerosis

The Multi-Ethnic Study of Atherosclerosis (MESA) was initiated in 2000 to investigate the pathogenesis of subclinical cardiovascular disease (CVD) in 6,814 men and women aged 45–84 years.[17] Participants were recruited from six U.S. communities and were free of CVD at baseline. Baseline measurements included CVD risk factors, demographic and psychosocial factors, life habits, and subclinical atherosclerosis. Selected risk factor and outcome variables were collected in the follow-up visits. All participants provided informed consent, and the study was approved by the Institutional Review Board at each site. Participants had a baseline examination (exam 1) in 2000–2002 and four additional follow-up examinations 18–24 months apart (exams 2–5). As exam 5 data were not available at the time of these analyses, only exams 1–4 were included. Obesity is an important CVD risk factor that is affected by genetics[30] and is modifiable by changing lifestyles (e.g., dietary intake, physical activity).[31,32] The goal of this analysis was to investigate how an overall exposure profile, including behavioral and psychosocial factors, modifies the genetic effects on body mass index (BMI). The analysis sample included 6429 MESA participants. Table 1 provides baseline demographic information for the study population. Approximately 39% of the cohort were Caucasian, 26% were African-American, 23% were Hispanic, and 12% were Chinese. The mean BMI at baseline was 28.8 kg/m^2. We analyzed G×E using the exposure profile groups generated by the aforementioned clustering methods. Below we describe the genetic variables and environmental exposure variables. Next, we applied Tukey’s one df, saturated interaction model, and additive main effects and multiplicative interaction model with M=1 (AMMI1) for the interactions.

Genes—Of 32 BMI-related SNPs according to genome-wide association study findings, we considered 27 SNPs that were available in all four ethnic groups with good imputation quality ($R^2 \geq 0.8$). Details on MESA genotyping and imputation to the HapMap 1+2 reference panel have been described previously.[33] We used the imputed genotypes with the highest imputed genotype probability. Three genotypes (homozygous for the non-risk allele, heterozygous, or homozygous for the risk allele) were considered for each SNP. We calculated the genetic risk score by summing BMI-increasing allele counts for the 27 SNPs. The genetic risk score was categorized into quintile categories to illustrate the use of Tukey’s and AMMI1 models.

Environmental Exposures—Table 2 lists 11 exposure variables in three domains: (1) dietary intake, (2) physical activity, and (3) psychosocial factors. The diet variables included total energy intake (kcal/day), percent calories from carbohydrate intake, protein intake, saturated fat intake, and trans fat intake. The physical activity variables included total intentional exercise (metabolic equivalent -minute/week) and moderate and vigorous physical activity (metabolic equivalent-minute/week). The psychosocial variables were trait anxiety, trait anger, chronic burden, and depressive symptoms (details in eAppendix).

Overall Exposure Profile Groups—To classify participants based on their overall exposure profile using the 11 exposure variables, we applied k-means, latent class analysis, classification and regression trees, and Bayesian Dirichlet process clustering. Total energy intake, intentional exercise, physical activities, and the four psychosocial variables were log-transformed to approximate normality. Since these methods are applicable to cross-sectional data, we conducted analysis with the 11 exposure variables at baseline as well as with subject-level averages for each exposure variable across MESA exams and found no appreciable difference in the exposure profile groups. We assumed constant exposure profile group membership (i.e. the probability of being in a latent class) across exams. Figure 1 shows the mean of each exposure variable corresponding to the clusters determined by k-means. Groups A, B, C, and D represent different dietary patterns. Group E has a physically active lifestyle, while Group F exhibits poor psychosocial health. Using latent class analysis, the best model was reached with an eight-class solution, but the model using six classes appeared to be more interpretable (eFigure 2 and eFigure 3 in eAppendix) and the Bayesian information criterion was close to the eight-cluster model. The classification and regression trees model demonstrated that percent calories from trans fats, chronic burden, and intentional exercise are significant predictors of BMI (Figure 2). To describe cluster characteristics in terms of the exposure variables using Bayesian Dirichlet process clustering, eFigure 4 displays the posterior cluster means. eFigure 5 shows the posterior distributions of cluster parameters for the representative clustering.

Genetic and Environmental Exposure Main Effects—In a pooled analysis of four ethnicities, we examined the genetic main effects and the environmental main effects on BMI using fixed-effects models with unstructured correlation structure for within-subject correlation (based on the smallest Akaike Information Criteria value). For environmental main effects, covariates included age, age square, gender, race/ethnicity, education, household income, and diagnosis of cancer. Participants selected their highest education level from eight categories that were collapsed into two: whether the subject attained a 4-year college degree. Participants identified their annual household income from 13 categories with different ranges of income ($0–$9,999, $10,000–$19,999, etc.) at MESA exams 1, 2, and 3. Continuous income in US dollars was assigned as the interval midpoint of the selected category. Except for age and income, all other covariates were collected at the baseline visit. For genetic main effects, covariates included age, age square, gender, and the first three genetic principal components for population stratification (together explained 96% of the total genetic variation).

Only rs2867125 (near *TMEM18*) and rs7359397 (near *SH2B1*, *APOB48R*, and *SULT1A2*) were associated with BMI after multiple testing correction (adjusted p-value = 0.05/27 = 0.0019). Except for trait anger and trait anxiety, all other environmental variables were associated with BMI. Table 3 shows the estimated main effects of the overall exposure profile clusters.

SNP/GRS × Overall Exposure Profile Interactions—We treated both G (three genotypes for each SNP) and E (represented by the exposure profile groups) as categorical variables and selected six SNP×E Profile interactions for subsequent tests based on a crude screening of all possible interactions. Table 4 shows the interaction test results between the six SNPs (and GRS) and overall exposure profile using Tukey's, AMMI1, and saturated interaction models. When using k-means to generate the overall exposure profile, AMMI1 and saturated models respectively detected two SNPs with modifying effects. rs1558902 near *FTO* and rs7359397 near *SH2B1* were found using the AMMI1 model (p=0.029 and p=0.047, respectively), and rs3817334 near *MTCH2* and again rs7359397 near *SH2B1* were found using the saturated interaction model (p=0.020 and p=0.043, respectively). Using classification and regression trees, rs543874 near *SEC16B* and again rs7359397 were found using both AMMI1 and saturated interaction models. Tukey's model also detected rs7359397. Using Bayesian Dirichlet process clustering, rs713586 near RBJ was detected using the saturated model (p=0.009), and rs3817334 near MTCH2 was found using AMMI1 model (p=0.039).

Genetic risk score was found to have a modifying effect on the association between the exposure clusters (defined by k-means) and BMI using all three models. Figure 3 shows different estimated effects of genetic risk score (comparing the 5^{th} and the 1^{st} quintile group) on BMI among six exposure profile groups, indicating a profound effect of genetic risk score on increased BMI for people with poor psychosocial health. Genetic risk score had a modifying effect on classification and regression tree groups using Tukey's model (p<0.0001) and the saturated model (p=0.014) and on Bayesian Dirichlet process clustering groups using the AMMI1 model (p=0.006) and saturated model (p=0.034). The results imply a Tukey's 1-df form of interaction between genetic risk score and overall exposure profile (grouped by k-means or classification and regression trees).

DISCUSSION

We introduced a novel way to integrate multiple environmental exposures in G×E analysis. A number of existing clustering and classification methods can be used to summarize information of multiple environmental exposures into one variable. This approach discovers the underlying grouping to guide hypothesis development. Moreover, using the summary variable for further G×E analysis can avoid repeated tests that could lead to reduced power after multiple comparison adjustment.

We described additive main effect and multiplicative interaction models for modeling G×E, as opposed to traditional saturated interaction modeling approach, to enhance test power by adopting a parsimonious interaction model. Testing the significance of each multiplicative term and selecting the optimal number of multiplicative terms in these models are natural follow-up questions. Many researchers have studied this problem primarily under balanced settings in yield trials.[34,35] Cross validation or parametric bootstrap can be applied to find the number of multiplicative interaction terms.[36,37] A comprehensive comparison of model-building strategies using the MESA data would be worthwhile but is beyond the scope of this paper.

The exposure profile group characteristics derived from k-means, latent class analysis, and Bayesian Dirichlet process clustering involve all 11 exposure variables under consideration, which can be difficult to interpret although we have described them qualitatively. The interpretation of classification and regression tree groups is straightforward because the algorithm identifies key factors and eliminates unimportant ones. We encourage researchers to know the strengths, limitations, and assumptions underlying each method to choose appropriate algorithms. G×G (or epistasis) was not considered here as they are typically of more interest to identify biological mechanisms. Our sample size also does not allow exploration of both G×G and G×E. From a technical point of view, creating clusters formed by multiple SNPs requires a different clustering treatment and entails an independent study. We restrict our attention to G×E with a focus on condensing exposure data.

A few limitations are worth noting. Since typical clustering approaches are not capable of handling repeated measurements, we computed the average exposure profile for each individual and performed cluster analysis. Not only does this simple strategy assume a time-invariant exposure profile but possibly leads to an inefficient model estimation. Future studies should explore the application of longitudinal clustering methods to identify time-changing exposure patterns and to discover time-varying G×E.

Cluster uncertainty is another critical issue. One can examine the uncertainty by bootstrapping the entire analytic process and create resampling-based sampling variance estimates. The PReMiuM package provides a principled and fully Bayesian way to account for clustering uncertainty in outcome model parameters. However, for easier interpretability of the exposure clusters as well as comparability with the other methods, we extracted the best clustering offered by the PReMiuM package and used it in the second step G×E model. This hybrid Bayes-frequentist approach is ad hoc and is a limitation of the method. This strategy was adopted because the G×E test using the additive main effects and multiplicative interaction model depends on the number of exposure clusters and we wanted to interpret the exposure profile in each cluster while performing the analysis. An integrated full Bayes implementation will be methodologically more appropriate though the hybrid approach has a practical appeal and is simple to understand.

An alternative is to marginalize over the clustering distribution or employ some form of Bayesian model averaging while reporting the final inference. We propose to pre-specify a set of potential exposure cluster numbers, fit individual clustering models, and then obtain Bayesian information criterion values. Then, we perform separate G×E tests for all exposure cluster models. Lastly, we summarize the G×E tests by weighing each test using the Bayesian information criterion value of that particular clustering model (see eAppendix eTable 3 for the G×E tests using BMA). However, Bayesian model averaging does not allow derivation of point estimates and confidence intervals for a practical interaction interpretation.

This paper demonstrates the use of clustering methods for translating multiple environmental exposures to one summary variable in G×E studies. One argument for defining categorical exposure sub-groups could be to better handle non-linearity, measurement error, and potential exposure–exposure interactions. It provides a useful characterization of G×E in terms of overall exposure profiles. More research in longitudinal G×E with multiple time-varying exposures is needed. With the advancement of powerful statistical tool and the availability of rich longitudinal data, we may identify time-dependent G×E effects and ultimately understand relationships among genes, environments, and complex diseases over different life stages.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Source of financial support: MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169 and CTSA UL1-RR-024156 and UL1-TR-001079. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. This project was conducted in part with funding from HL101161 from NHLBI and 2P60MD002249 from NIMHD. Research of BM and YK was supported by NIH/NIEHS grant ES-20811 and NSF grant DMS-1406712. The National Heart, Lung and Blood Institute (NHLBI) is the primary MESA funding source. The MESA Project Officer (at NHLBI) is represented on the MESA Steering Committee and the Executive Committee which oversees design and data collection decisions and reviews MESA manuscripts before they are submitted for publication. The funder was not involved in the preparation of the manuscript or in the final decision to publish.

References

1. Fan R, Albert PS, Schisterman EF. A discussion of gene-gene and gene-environment interactions and longitudinal genetic analysis of complex traits. Stat Med. 2012; 31:2565–8. [PubMed: 22969024]
2. Zhang A, Park SK, Wright RO, et al. HFE H63D polymorphism as a modifier of the effect of cumulative lead exposure on pulse pressure: the Normative Aging Study. Environ Health Perspect. 2010; 118:1261–6. [PubMed: 20478760]
3. Julvez J, Grandjean P. Genetic susceptibility to methylmercury developmental neurotoxicity matters. Front Genet. 2013; 4:278. [PubMed: 24379825]
4. Zou F, Huang H, Lee S, Hoeschele I. Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene-environment interaction. Genetics. 2010; 186:385–94. [PubMed: 20551445]
5. Greenland S. Methods for epidemiologic analyses of multiple exposures - a review and comparative-study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med. 1993; 12:717–36. [PubMed: 8516590]
6. Loska K, Wiechula D. Application of principal component analysis for the estimation of source of heavy metal contamination in surface sediments from the Rybnik Reservoir. Chemosphere. 2003; 51:723–33. [PubMed: 12668031]
7. Enoch MA, Baghal B, Yuan Q, Goldman D. A factor analysis of global GABAergic gene expression in human brain identifies specificity in response to chronic alcohol and cocaine exposure. PLoS One. 2013; 8:e64014. [PubMed: 23717525]
8. Moreno-Macias H, Romieu I, London SJ, Laird NM. Gene-environment interaction tests for family studies with quantitative phenotypes: A review and extension to longitudinal measures. Hum Genomics. 2010; 4:302–26. [PubMed: 20650819]
9. Rosario AS, Wellmann J, Heid IM, Wichmann HE. Radon epidemiology: continuous and categorical trend estimators when the exposure distribution is skewed and outliers may be present. J Toxicol Environ Health A. 2006; 69:681–700. [PubMed: 16608833]
10. Siahpush SH, Vaughan TL, Lampe JN, et al. Longitudinal study of insulin-like growth factor, insulin-like growth factor binding protein-3, and their polymorphisms: risk of neoplastic progression in Barrett's esophagus. Cancer Epidemiol Biomarkers Prev. 2007; 16:2387–95. [PubMed: 18006928]
11. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet. 2006; 79:1002–16. [PubMed: 17186459]
12. Mukherjee B, Ko YA, VanderWeele T, Roy A, Park SK, Chen JB. Principal interactions analysis for repeated measures data: application to gene-gene and gene-environment interactions. Stat Med. 2012; 31:2531–51. [PubMed: 22415818]
13. Ko YA, Saha-Chaudhuri P, Park SK, Vokonas PS, Mukherjee B. Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data. Genet Epidemiol. 2013; 91–37:581.
14. Gauch HG. Model selection and validation for yield trials with interaction. Biometrics. 1988; 44:705–15.
15. Gollob HF. A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika. 1968; 33:73–115. [PubMed: 5239571]
16. Mandel J. New analysis of variance model for non-additive data. Technometrics. 1971; 13:1–18.
17. Bild DE, Bluemke DA, Burke GL, et al. Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol. 2002; 156:871–81. [PubMed: 12397006]
18. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010; 31:651–66.
19. Jain, AK.; Dubes, RC. Algorithms for clustering data. Englewood Cliffs, N.J: Prentice Hall; 1988.
20. Lazarsfeld, P.; Henry, N. Latent structure analysis. Boston: Houghton Mifflin; 1968.
21. Fraley C, Raftery AE. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. J Classif. 2003; 20:263–86.

22. Breiman, LFJ.; Olshen, RA.; Stone, CJ. Classification and regression trees. Belmont: Wadsworth; 1984.

23. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National survey of children's health. Biostatistics. 2010; 11:484–98. [PubMed: 20350957]

24. Escobar MD, West M. Bayesian Density-Estimation and Inference Using Mixtures. J Am Stat Assoc. 1995; 90:577–88.

25. Muller P, Rosner GL. A Bayesian population model with hierarchical mixture priors applied to blood count data. J Am Stat Assoc. 1997; 92:1279–92.

26. Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. J Stat Softw. 2015; 64:1–30. [PubMed: 27307779]

27. Strobl C, Boulesteix AL, Augustin T. Unbiased split selection for classification trees based on the Gini Index. Comput Stat Data An. 2007; 52:483–501.

28. Tukey JW. One Degree of Freedom for Non-Additivity. Biometrics. 1949; 5:232–42.

29. Marasinghe MG, Johnson DE. A Test of incomplete additivity in the multiplicative interaction-model. J Am Stat Assoc. 1982; 77:869–77.

30. Maes HHM, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. Behav Genet. 1997; 27:325–51. [PubMed: 9519560]

31. Wadden TA, Webb VL, Moran CH, Bailer BA. Lifestyle modification for obesity new developments in diet, physical activity, and bevaior therapy. Circulation. 2012; 125:1157–70. [PubMed: 22392863]

32. Onyike CU, Crum RM, Lee HB, Lyketsos CG, Eaton WW. Is obesity associated with major depression? Results from the Third National Health and Nutrition Examination Survey. Am J Epidemiol. 2003; 158:1139–47. [PubMed: 14652298]

33. Rasmussen-Torvik LJ, Guo X, Bowden DW, et al. Fasting glucose GWAS candidate region analysis across ethnic groups in the Multiethnic Study of Atherosclerosis (MESA). Genet Epidemiol. 2012; 36:384–91. [PubMed: 22508271]

34. Pawar Y, Patil HT, Patil HS. AMMI analysis for grain yield stability of pearl millet (Pennisetum glaucum L) genotypes Indian. J Genet Pl Br. 2012; 72:79–82.

35. Rea R, De Sousa-Vieira O, Diaz A, et al. Genotype-environment interaction in sugarcane by AMMI and site regression models in Venezuela. Rev Fac Agron Luz. 2014; 31:362–76.

36. Dias CTD, Krzanowski WJ. Model selection and cross validation in additive main effect and multiplicative interaction models. Crop Sci. 2003; 43:865–73.

37. Forkman J, Piepho HP. Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. Biometrics. 2014; 70:639–47. [PubMed: 24588726]

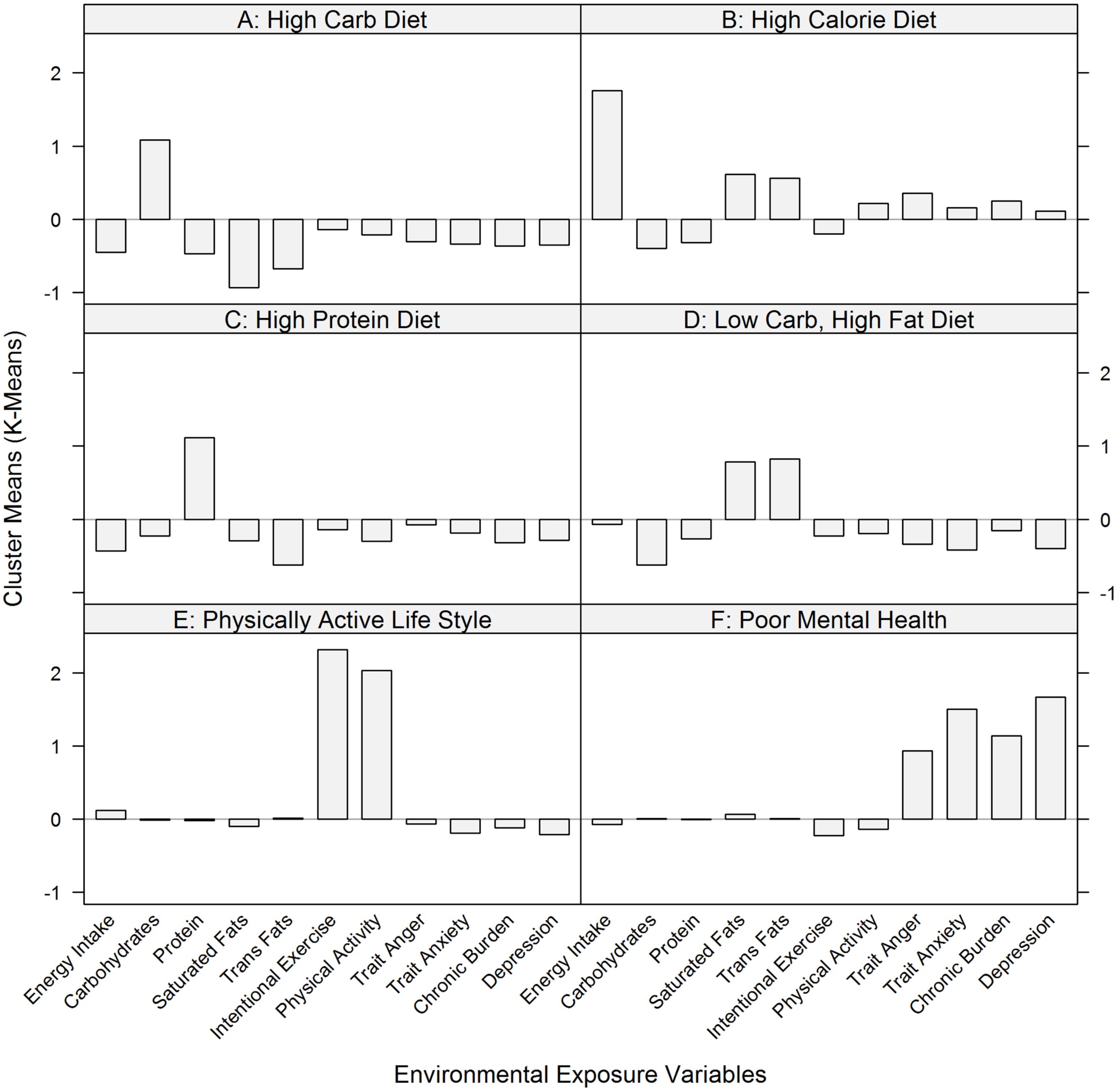

Figure 1.

Cluster means of the 11 (standardized) environmental exposure variables using k-means in the MESA data

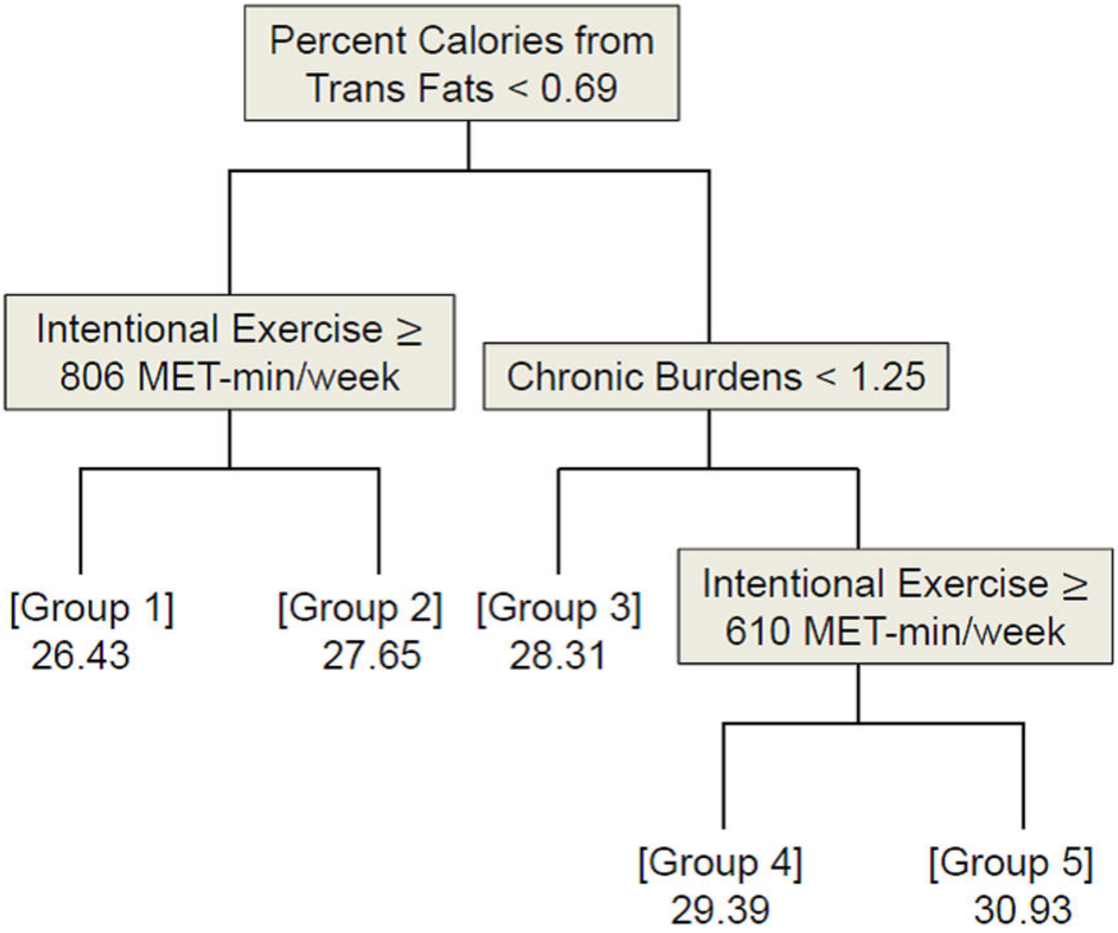
Percent Calories from Trans Fats < 0.69
Intentional Exercise ≥ 806 MET-min/week
Chronic Burdens < 1.25
[Group 1] 26.43
[Group 2] 27.65
[Group 3] 28.31
Intentional Exercise ≥ 610 MET-min/week
[Group 4] 29.39
[Group 5] 30.93

Figure 2.

Grouping criteria of classification and regression tree (CART) analysis results. Mean BMI value is shown for each group.

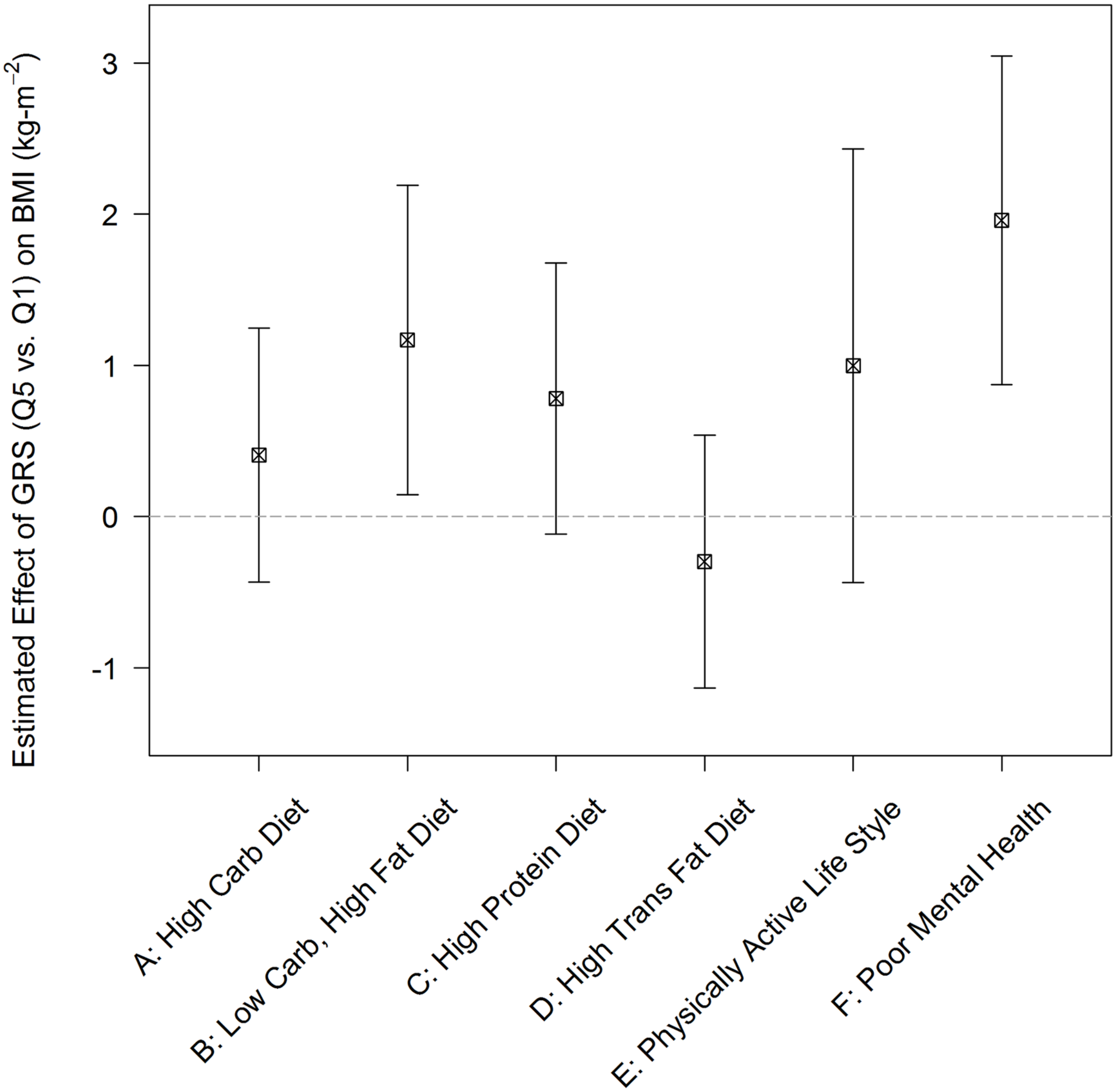

Figure 3.

Estimated effects of the genetic risk score (GRS) on BMI (and the corresponding 95% confidence intervals) comparing the fifth (Q5) to the first GRS quintile group (Q1) for the six exposure profile groups obtained by k-means.

Table 1

Baseline characteristics of the study participants in the Multi-Ethnic Study of Atherosclerosis.

	CAU (N=2527)		**CHN (N=775)**		**AFA (N=1677)**		**HIS (N=1450)**		**All (N=6429)**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Gender (female)	1286	52	389	51	842	55	722	51	3239	52
Education										
Less than high school degree	118	5	190	25	167	11	629	45	1104	18
High school degree or above, without bachelor degree	1108	45	278	36	837	54	646	46	2869	46
Bachelor degree or above	1235	50	301	39	538	35	139	10	2213	36
Total annual gross family income										
Less than $20,000	270	11	323	42	335	22	562	40	1490	24
$20,000 or more but less than $50,000	794	32	228	30	633	41	602	43	2257	37
$50,000 or more	1397	57	218	28	574	37	250	18	2439	39
Diagnosis of cancer	318	13	19	2	103	6	71	5	511	8
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age (years)	66.6	10.3	64.7	10.4	64.2	9.9	62.4	10.3	64.3	10.3
Body mass index (kg/m^2)	28.4	5.4	23.7	3.2	30.4	6.1	29.5	5.2	28.8	5.7
GRS (count of BMI-increasing alleles)	23.5	3.2	23.4	3.3	23.8	2.9	23.3	3.2	23.7	3.1

CAU = Caucasian, CHN = Chinese, AFA = African American, HIS = Hispanic, GRS = genetic risk score, SD = standard deviation.

Table 2

Baseline summary of environmental exposure variables for the study participants in MESA

	CAU		CHN		AFA		HIS		All	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Total energy intake (kcal/day)	1514	755	1055	473	1599	955	1605	856	1505	836
Percent calories from carbohydrate intake	51.7	9.0	56.2	7.7	53.3	9.1	55.1	7.9	53.9	8.6
Percent calories from protein intake	15.3	3.3	17.9	3.3	15.3	3.2	16.0	3.2	15.9	3.4
Percent calories from saturated fat intake	11.2	3.4	7.8	2.4	10.2	2.9	10.4	3.1	10.2	3.2
Percent calories from trans fat intake	0.9	0.3	0.5	0.2	1.0	0.4	0.7	0.3	0.8	0.4
Total intentional exercise (MET-min/wk)	1467	2509	1023	1383	1607	2615	1187	1973	1349	2252
Physical activity[a] (MET-min/wk)	6268	6217	3595	3734	6237	6546	5629	6184	5695	6099
Trait anger	14.5	3.3	14.7	3.5	14.2	3.6	15.0	4.4	14.6	3.9
Trait anxiety	17.0	4.7	16.2	4.7	15.7	4.6	16.2	4.8	16.3	4.7
Chronic burden	1.3	1.2	0.8	1.1	1.4	1.3	1.2	1.2	1.2	1.2
Depressive symptoms (CESD)	8.4	7.3	6.2	6.8	8.5	8.1	10.0	9.2	8.7	8.2

CAU = Caucasian, CHN = Chinese, AFA = African American, HIS = Hispanic, CESD = Center for Epidemiologic Studies Depression Scale, MET = metabolic equivalent, SD = standard deviation

[a] Moderate and vigorous physical activity

Table 3

Estimates of the exposure cluster main effects on BMI adjusted for age, age^2, gender, race, education, income, and diagnosis of cancer

Effect	Estimate	95% Confidence Limits
K-means: Group B vs. Group A	1.9	(1.4, 2.3)
K-means: Group C vs. Group A	0.76	(0.26, 1.1)
K-means: Group D vs. Group A	1.2	(0.82, 1.6)
K-means: Group E vs. Group A	0.15	(−0.41, 0.70)
K-means: Group F vs. Group A	1.2	(0.77, 1.7)
LCA: Group B vs. Group A	0.06	(−0.35, 0.47)
LCA: Group C vs. Group A	1.2	(0.70, 1.7)
LCA: Group D vs. Group A	0.83	(0.44, 1.2)
LCA: Group E vs. Group A	0.02	(−0.39, 0.4)
LCA: Group F vs. Group A	1.6	(1.1, 2.0)
CART: Group 2 vs. Group 1	1.2	(0.78, 1.6)
CART: Group 3 vs. Group 1	1.1	(0.86, 1.5)
CART: Group 4 vs. Group 1	2.0	(1.6, 2.4)
CART: Group 5 vs. Group 1	3.3	(2.8, 3.7)
BDPC: Group 2 vs. Group 1	1.7	(1.4, 2.1)
BDPC: Group 3 vs. Group 1	−0.50	(−1.0, 0.02)
BDPC: Group 4 vs. Group 1	0.79	(0.29, 1.3)
BDPC: Group 5 vs. Group 1	−1.5	(−1.8, −1.1)

See figures for the characteristics of each group classified by k-means, latent class analysis (LCA), and classification and regression tree (CART), and Bayesian Dirichlet Process clustering (BDPC).

Table 4

Test results (p-values) of interaction between overall exposure profile groups (using k-means, LCA, CART, and BDPC) and BMI-related SNPs using Tukey's single degree of freedom, AMMI1, and saturated (SAT) interaction models in the MESA data

				K-means			**LCA**			**CART**			**BDPC**	
SNP	**SNP ID**	**Gene**	**Tukey**	**AMMI1**	**SAT**	**Tukey**	**AMMI**	**SAT**	**Tukey**	**AMMI1**	**SAT**	**Tukey**	**AMMI1**	**SAT**
2	rs543874	SEC16B	0.509	0.628	0.516	0.424	0.590	0.552	0.987	0.018	0.007	0.209	0.364	0.304
6	rs713586	RBJ, ADCY3, POMC	0.143	0.761	0.836	0.798	0.278	0.155	0.100	0.200	0.247	0.067	0.080	0.009
16	rs3817334	MTCH2, NDUFS3	0.799	0.051	0.020	0.820	0.128	0.228	0.443	0.359	0.476	0.832	0.039	0.332
19	rs1015033	NRXN3	0.443	0.999	1.000	0.583	0.822	0.831	0.548	0.829	0.907	0.186	0.354	0.259
22	rs1558902	FTO	0.296	0.029	0.056	0.534	0.762	0.841	0.558	0.456	0.612	0.106	0.136	0.199
23	rs7359397	SH2B1, APOB48R	0.111	0.047	0.043	0.769	0.346	0.307	0.009	0.006	0.005	0.900	0.476	0.630
GRS[a]			0.033	0.022	0.003	0.367	0.220	0.132	9×10^{-5}	0.136	0.014	0.277	0.006	0.034

[a] The genetic risk score (GRS) was categorized into 5 groups based on quintiles. Covariates were age, age^2, gender, race, education, income, diagnosis of cancer, and the first three principal components for population stratification