Environment International 118 (2018) 146–153

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Toward a panoramic perspective of the association between environmental factors and cardiovascular disease: An environment-wide association study from National Health and Nutrition Examination Survey 1999–2014

Xiaodong Zhuang[a,b,1], Yue Guo[a,b,1], Ao Ni[c,1], Daya Yang[a,b], Lizhen Liao[d], Shaozhao Zhang[a,b], Huimin Zhou[a,b], Xiuting Sun[a,b], Lichun Wang[a,b], Xueqin Wang[c,e,*], Xinxue Liao[a,b,**]

[a] *Cardiology Department, The First Affiliated Hospital of Sun Yat-Sen University, China*
[b] *Key Laboratory on Assisted Circulation, Ministry of Health, China*
[c] *Department of Statistical Science, School of Mathematics and Computational Science, Sun Yat-Sen University, China*
[d] *Department of Health, Guangdong Pharmaceutical University, Guangzhou Higher Education Mega Center, China*
[e] *Joint Institute of Engineering, Sun Yat-Sen University-Carnegie Mellon University, China*

ARTICLE INFO

Handling Editor: Olga-Ioanna Kalantzi

Keywords:
Cardiovascular disease
Environment-wide association study
Random forest
False discovery rate
Receiver operating characteristic

ABSTRACT

Objectives: An environment-wide association study (EWAS) may be useful to comprehensively test and validate associations between environmental factors and cardiovascular disease (CVD) in an unbiased manner.
Approach and results: Data from National Health and Nutrition Examination Survey (1999–2014) were randomly 50:50 spilt into training set and testing set. CVD was ascertained by a self-reported diagnosis of myocardial infarction, coronary heart disease or stroke. We performed multiple linear regression analyses associating 203 environmental factors and 132 clinical phenotypes with CVD in training set (false discovery rate < 5%) and significant factors were validated in the testing set ($P < 0.05$). Random forest (RF) model was used for multicollinearity elimination and variable importance ranking. Discriminative power of factors for CVD was calculated by area under the receiver operating characteristic (AUROC). Overall, 43,568 participants with 4084 (9.4%) CVD were included. After adjusting for age, sex, race, body mass index, blood pressure and socio-economic level, we identified 5 environmental variables and 19 clinical phenotypes associated with CVD in training and testing dataset. Top five factors in RF importance ranking were: waist, glucose, uric acid, and red cell distribution width and glycated hemoglobin. AUROC of the RF model was 0.816 (top 5 factors) and 0.819 (full model). Sensitivity analyses reveal no specific moderators of the associations.
Conclusion: Our systematic evaluation provides new knowledge on the complex array of environmental correlates of CVD. These identified correlates may serve as a complementary approach to CVD risk assessment. Our findings need to be probed in further observational and interventional studies.

1. Introduction

Cardiovascular diseases (CVD), a leading cause of mortality worldwide, is a heterogeneous and multifactorial disease, influenced by multiple genetic and environmental factors (Krittanawong and Kitai, 2017; Pasipoularides, 2015). Environmental issues, such as chemical toxicants, pollutants, allergens, bacterial/viral organisms, and nutrients, constitute an essential and often overlooked component (Munzel and Daiber, 2018). However, most of the documented researches often test associations of single or several environmental variables with CVD, potentially leading to incomplete understanding or misleading notions about possible contributors. Hence, it is important to identify potential environmental factors that may influence the progression of CVD with a panoramic perspective.

We borrowed the genome-wide association study (GWAS) methodology creating a model environment-wide association study (EWAS), to search for environmental factors associated with diseases on a broad scale (McGinnis et al., 2016; Zhuang et al., 2018). EWAS examined variables with a systematic approach, thus avoiding selective reporting bias and controlling for the rate of false positives (Patel et al., 2016). Previous EWAS have scanned for associations between environmental exposure and behavioral factors putatively correlated with diseases,

[*] Correspondence to: X. Wang, Department of Statistical Science, School of Mathematics and Computational Science, Sun Yat-Sen University, China.
[**] Correspondence to: X. Liao, The First Affiliated Hospital of Sun Yat-Sen University, China.
E-mail addresses: wangxq@mail.sysu.edu.cn (X. Wang), liaoxinx@mail.sysu.edu.cn (X. Liao).
[1] Authors contribute equally to the study: Xiaodong Zhuang, Yue Guo; Ao Ni.

https://doi.org/10.1016/j.envint.2018.05.046
Received 19 February 2018; Received in revised form 25 May 2018; Accepted 26 May 2018

such as type 2 diabetes, peripheral arterial disease, and all-cause mortality (Zhuang et al., 2018; Patel et al., 2013; Hall et al., 2014).

We extend the EWAS approach to evaluate multiple associations between a wide range of environmental exposures and CVD using the dataset from the US National Health and Nutrition Examination Survey (NHANES), a nationally representative, cross-sectional biannual health survey (Patel et al., 2016). Specific environmental attributes and clinical phenotypes are assayed in NHANES. Such an approach can prioritize environmental factors for future investigation, providing us insight in regards to CVD etiology and prognosis prediction.

2. Methods

The schematic representation of our analysis methodology is summarized in Supplement 1.

2.1. Study population

The NHANES is a publicly available dataset made available by the Centers for Disease Control and Prevention (CDC) and National Centers for Health Statistics (Patel et al., 2016). Protocol approval and written informed consent were obtained by the National Center for Health Statistics Institutional Review Board for participants > 18 years of age and from the guardians of participants < 18. All methods were carried out by the approved guidelines. All survey and consent documents for NHANES were approved by the CDC Institutional Review Board.

We used data from eight cross-sectional surveys (biannual from 1999 to 2014). This cross-sectional dataset is comprised of health questionnaire, laboratory (i.e., urinary phthalates, blood lead, blood cadmium, urinary mercury), and clinical data using a multistage probability sampling design. Data were collected through in-person interviews, physical measurement at mobile examination centers and laboratory samples. Self-reported data were also collected for supplement intake, diabetes mellitus, or cardiovascular disease status, family history of hypertension, and fitness level coded as metabolic equivalent of task (MET).

2.2. Cardiovascular diseases definition

Definition of CVD was ascertained by self-reported questionnaires: “Has a doctor or other health professional ever told you that you had a heart attack (also called myocardial infarction)?” or “Has a doctor or other health professional ever told you that you had a coronary artery disease?” or “Has a doctor or other health professional ever told you that you had a stroke?” Answering yes to either question was coded positive for CVD.

2.3. Environmental exposures in the EWAS

There were a total of 335 factors, including 203 environmental factors and 132 clinical phenotypes, in our EWAS (Supplement 2). Environmental factors comprised of 7 bacterial infections, 29 furans, 35 heavy metals, 11 hydrocarbons, 23 PCBs, 20 pesticides, 11 phenols, 16 phthalates, 12 polyflourochemicals, 12 viral infection, and 34 volatile compounds. Other clinical phenotypes comprised of 44 biochemistry, 11 blood pressure, 19 blood routine, 8 body measure, 22 DXA, 15 nutrients, and 13 spirometry factors. Different environmental factors were measured in varying numbers of participants, ranging from 533 to 16,721 individuals over the different environmental factors. Based on the power calculation result for general linear model we did in advance, a minimum sample size of 500 in our study can reach at least 97% power. So we include factors with sample size larger than 500 in the analysis to insure enough statistical power and maximal number of variables. We removed factors that targeted a subset of the population, such as the test for Trichomonas vaginalis, an infectious pathogen found primarily in women. We also omitted the factors that varied little across individuals and those that had a majority (90%) of the observations below a detection limit threshold as defined by in the NHANES codebook.

2.4. Statistical analyses

All analyses were performed in R version 3.3.0 (The R Foundation for Statistical Computing, www.R-project.org). All variables in our study were either continuous or discrete. We log transform the continuous variables that had right-skewed distribution; then we z-standardized all continuous variables as previously described (Zhuang et al., 2018).

EWAS was performed as previously described (Zhuang et al., 2018; Patel et al., 2013). Briefly, we conducted regression analysis and accounted for clusters pseudo-strata, pseudo-sampling units, and participant weights to accommodate the complex sampling of the data. We did a random 50:50 spilt of the dataset into training set and testing set to validate our result within the dataset. We associate each of 342 environmental factors with CVD using survey-weighted logistic regression model, adjusting for age, sex, ethnicity, and body measure index (BMI), systolic blood pressure (SBP) and socio-economic status (SES) level. In the training set, we use Benjamin-Hochberg procedure to control the false discovery rate (FDR) at level 5%. We deemed a factor tentatively validated if it had achieved FDR < 5% significance in the training set and reached nominal statistical significance in the testing set (P value < 0.05). For tentatively validated factors, we computed its' odds ratio (OR) and 95% confidence interval (CI). We further analyzed the identified factors using a random forest model for multicollinearity elimination and ranked the identified factors based on variable importance scores. Additionally, the discriminative power of the top five key factors and their combination for CVD was calculated by receiver operating characteristic (ROC) curves.

We conducted the subsequent analytic tests for the validity and sensitivity of our final estimates. Firstly, we assessed the Spearman correlations among all validated factors and visualized these variables in a heat map. The hierarchical clustering algorithm was used to arrange these factors in the heat map. The more significant the correlation between a pair of variables, the closer in proximity they appear in the heat map. Secondly, we assessed the Spearman correlations among all validated factors. Strong association with spearman's P value < 0.05 and the absolute value of Spearman's correlation coefficient > 2.6 was displayed in a network graph. Moreover, we conducted a sensitivity analysis to investigate the potential influence of three factors including age, gender, cholesterol-lowing medicine, and different definition of CVD subtypes. Finally, we conducted additional meta-analyses of all identified factors per survey using the I^2 statistic to evaluate the heterogeneity between survey year.

3. Results

Overall 43,568 participants were included in the EWAS, with 4084 (9.4%) participants defined as CVD, the prevalence of which was 3.9% (stroke), 4.2% (CAD), and 4.5% (MI). Table 1 describes the baseline and demographic characteristics of people with or without CVD in the survey. There were significant differences with baseline factors such as age, sex, race, SES, and comorbidity between the two groups. CVD occurred in higher age, more male, lower SES level and higher BMI in all cohorts (P < 0.001, two-sided t-test). Furthermore, there were significantly more patients with congestive heart failure, as well as cigarette smoking, in the CVD group for all cohorts.

3.1. Systematic scan of environmental factors associated with cardiovascular diseases

After adjusting for age, sex, race, BMI, SBP and SES level, the distribution of P values of association for each environmental factor and