Original Article

Generalized Functional Linear Models: Efficient Modeling for High-dimensional Correlated Mixture Exposures

Bingsong Zhang[1], Haibin Yu[1], Xin Peng[1], Haiyi Yan[1], Siran Li[1], Shutong Luo[2], Renhuizi Wei[3], Zhujiang Zhou[1], Yalin Kuang[1], Yihuan Zheng[1], Chulan Ou[1], Linhua Liu[4,#], Yuehua Hu[5,#], and Jindong Ni[6,7,#]

1. Department of Epidemiology and Biostatistics, School of Public Health, Guangdong Medical University, Dongguan 523808, Guangdong, China; 2. Herbert Wertheim School of Public Health, Warrant College, University of California San Diago 92037, U.S.A; 3. Office of Quality Management, Hospital of Huangjiang Dongguan, Dongguan 523750, Guangdong, China; 4. Dongguan Key Laboratory of Environmental Medicine, School of Public Health, Guangdong Medical University, Dongguan 523808, Guangdong, China; 5. Office of Epidemiology, Chinese Center for Disease Control and Prevention, Beijing 102206, China; 6. Precision Key Laboratory of Public Health, School of Public Health, Guangdong Medical University, Dongguan 523808, Guangdong, China; 7. Maternal and Child Research Institute, Shunde Women and Children's Hospital, Guangdong Medical University, Foshan 528300, Guangdong, China

Abstract

Objective Humans are exposed to complex mixtures of environmental chemicals and other factors that can affect their health. Analysis of these mixture exposures presents several key challenges for environmental epidemiology and risk assessment, including high dimensionality, correlated exposure, and subtle individual effects.

Methods We proposed a novel statistical approach, the generalized functional linear model (GFLM), to analyze the health effects of exposure mixtures. GFLM treats the effect of mixture exposures as a smooth function by reordering exposures based on specific mechanisms and capturing internal correlations to provide a meaningful estimation and interpretation. The robustness and efficiency was evaluated under various scenarios through extensive simulation studies.

Results We applied the GFLM to two datasets from the National Health and Nutrition Examination Survey (NHANES). In the first application, we examined the effects of 37 nutrients on BMI (2011–2016 cycles). The GFLM identified a significant mixture effect, with fiber and fat emerging as the nutrients with the greatest negative and positive effects on BMI, respectively. For the second application, we investigated the association between four pre- and perfluoroalkyl substances (PFAS) and gout risk (2007–2018 cycles). Unlike traditional methods, the GFLM indicated no significant association, demonstrating its robustness to multicollinearity.

Conclusion GFLM framework is a powerful tool for mixture exposure analysis, offering improved handling of correlated exposures and interpretable results. It demonstrates robust performance across various scenarios and real-world applications, advancing our understanding of complex environmental exposures and their health impacts on environmental epidemiology and toxicology.

Key words: Mixture exposure modeling; Functional data analysis; High-dimensional data; Correlated exposures; Environmental epidemiology

[#]Correspondence should be addressed to Jindong Ni, Professor, PhD, Tel: 86-15817668208, E-mail: nijd-gw@gdmu.edu.cn; Yuehua Hu, Professor, PhD, Tel: 86-15811185438, E-mail: huyh@chinacdc.cn; Linhua Liu, Professor, PhD, Tel: 86-13620060551, E-mail: linhua-liu@gdmu.edu.cn

Biographical notes of the first authors: Bingsong Zhang, PhD, Assistant Professor, majoring in functional data analysis and causal inference, E-mail: zhangbingsong@gdmu.edu.cn; Haibin Yu, PhD, Associated Professor, majoring in data mining, E-mail: hby616688@gdmu.edu.cn

Biomed Environ Sci, 2025; 38(8): 961-976 *doi: 10.3967/bes2025.024* *ISSN: 0895-3988*
www.besjournal.com (full text) *CN: 11-2816/Q*

INTRODUCTION

Humans are exposed to complex mixtures of environmental chemicals and other factors that can affect their health. The importance of studying mixed exposures has been increasingly recognized in recent years[1,2]. For example, a study on prenatal exposure to multiple endocrine-disrupting chemicals revealed associations with altered cognitive function in children[3], whereas another investigation revealed that a mixture of air pollutants was linked to increased cardiovascular disease risk[4]. Analyzing the health effects of these multi-pollutant or multi-exposure mixtures presents several key challenges in environmental epidemiology and risk assessment. First, the mixture exposure data are often high-dimensional, with the number of exposures approaching or exceeding the sample size. Second, exposures within a mixture are frequently correlated with each other, leading to issues with multicollinearity in traditional regression models. Third, individual exposures often exert subtle effects that may not be statistically significant on their own. However, when these exposures accumulate in large numbers, they can potentially have a significant effect on health outcomes[5,6]. The primary goal of mixture exposure studies is to test the overall effect of combined exposure on health outcomes while adjusting for relevant covariates. In many cases, researchers aim to determine the relative contributions of individual exposures within a mixture[7,8].

Several statistical approaches have been developed to address the challenges in analyzing exposure mixtures. A weighted quantile sum (WQS) regression constructs a weighted index of the mixture components to estimate the overall mixture effect while identifying the relative importance of individual exposures[9]. WQS is computationally efficient and provides easily interpretable results; however, it assumes that all exposures have effects in the same direction, which may not always be realistic. An extension, known as the two-index WQS, allows for exposures with effects in opposite directions by constructing two separate indices[10]. This approach provides more flexibility, but may still struggle with highly correlated exposures. Quantile-based g-computation further extends the WQS framework by allowing exposure to effects in opposite directions and providing unbiased estimates of the overall mixture effect[11]. However, it is based on a generalized linear regression framework and may not fully capture the complex nonlinear exposure-response relationships. Bayesian kernel machine regression (BKMR) can handle high-dimensional data and provides a framework for variable selection. However, its results can be difficult to interpret, and the method is computationally intensive for large datasets[12].

In this study, we propose a novel statistical approach for analyzing the health effects of exposure mixtures using functional data analysis (FDA) techniques. This method treats the effect of mixture exposure as a smooth function and captures internal correlations to provide meaningful estimation and interpretation. This idea is inspired by approaches used in analyzing single nucleotide polymorphism (SNP) data, which share similarities with mixture exposure data in that individual components may have subtle effects, but their combination can significantly impact outcomes[13,14]. Our method leverages the strengths of the FDA to handle the high-dimensional and correlation structure inherent in mixture exposure data while allowing for flexible modeling of nonlinear mixture exposure–response relationships. We developed this approach, conducted extensive simulation studies to evaluate its robustness and efficiency, and applied it to analyze two datasets from the National Health and Nutrition Examination Survey (NHANES). This new framework provides a powerful tool for mixture exposure analysis, offering an improved handling of correlated exposures and interpretable results. It has the potential to advance our understanding of complex environmental exposures and their health impacts, with applications in various fields of environmental epidemiology and toxicology.

METHODS

Generalized Functional Linear Model (GFLM)

Model Consider n individuals with measured data for a mixture of m exposed substances. We assume that the m exposures are sorted in a random order $1 < \cdots < m$ in the collected dataset. For the i th individual, let y_i denote the response of interest,

which can be either quantitative or dichotomous, $E_i = (e_{i1}, \cdots, e_{im})^T$ denote the measured level of m exposures, and $Z_i = (z_{i1}, \cdots, z_{ic})^T$ denote the covariates. To relate the exposure mixtures to the response, while adjusting for covariates, we constructed the following model:

$$g(E(y_i)) = \alpha_0 + Z_i^T\alpha + \sum_{j=1}^{m} e_{ij}\beta(x_j), \quad (1)$$

in which $g(\cdot)$ is an identical link for the quantitative response and logit link for the dichotomous response, α_0 is the overall mean, α is an $c \times 1$ vector of regression coefficients for covariates, and $\beta(x_j)(j = 1, \cdots, m)$ is the exposure effect function (EEF) of position x_j. Under the functional regression framework, $\beta(x)$ is assumed to be a smooth function to reduce the number of coefficients needed to be estimated, whereas, in practice, it is discrete and assumed to be measurable at position x.

Estimation of EEF Generally, in research studies, there is no inherent order relationship among exposures. These are typically recorded randomly in a dataset, meaning that any two adjacent exposures may be positively, negatively, or uncorrelated. However, Model (1) assumes that the effect of exposure is a smooth function, implying that the arrangement of exposures cannot be random. Consider the following scenario: Suppose that e_{j_1} and e_{j_2} are two adjacent exposures in the dataset, where e_{j_1} is positively correlated with the outcome, but where e_{j_2} is negatively correlated. In this case, a true effect function needs to transition from a positive to a negative value within a short interval. This is unrealistic for achieving an estimated smooth function without overfitting. Therefore, exposures should be sorted according to a mechanism Δ to obtain a well-representative effect function $\beta(\cdot)$. To mitigate the impact of varying measurement units on effect size estimation, all exposure measurements were transformed into quartiles or decimals. Let $Q_i(= 0, 1, 2, 3)$ denote the quartile version of the exposure data E_i, where $(\Delta_1, \cdots, \Delta_m)$ is the new order of the exposure; then, we can use $Q_i = \left(q_{i\Delta_1}, \cdots, q_{i\Delta_m}\right)^T$ to substitute E_i as

$$g(\mathrm{E}(y_i)) = \alpha_0 + Z_i^T\alpha + \sum_{j=\Delta_1}^{\Delta_m} q_{ij}\beta(x_j). \quad (2)$$

To estimate the EEF $\beta(x)$, we can use an ordinary linear square (OLS) smoother[15]. Under the OLS framework, let $\psi_k(t), (k = 1, \cdots, K, K < m)$ be a series of basis functions. Two types of commonly used basis functions are (1) the B-spline basis $B_k(t), k = 1, \cdots, K$ and (2) the Fourier basis $F_0(t) = 1, F_{2r-1}(t) = \sin(2\pi rt)$, and $F_{2r}(t) = \cos(2\pi rt), r = 1, \cdots, (K-1)/2$. For the Fourier basis, K is a positive odd integer[15-19]. Then, $\beta(x)$ can be expanded as

$$\beta(x) = (\psi_1(x), \cdots, \psi_K(x))(\beta_1, \cdots, \beta_K)^T := \Psi\mathrm{B}, \quad (3)$$

where Ψ is an $m \times K$ matrix and where $\mathrm{B} = (\beta_1, \cdots, \beta_K)^T$ is a vector of coefficients. Thus, Model (1) can be rewritten as

$$g(\mathrm{E}(y_i)) = \alpha_0 + Z_i^T\alpha + \sum_{j=\Delta_1}^{\Delta_m} q_{ij}\beta(j) = \alpha_0 + Z_i^T\alpha + Q_i^T\Psi\mathrm{B}, \quad (4)$$

in which $\mathrm{W_i} = Q_i^T\Psi$.

Note that

$$g(\mathrm{E}(y_i|Q_i = 1)) - g(\mathrm{E}(y_i|Q_i = 0)) = 1^T\Psi\mathrm{B}, \quad (5)$$

where 1 is an $m \times 1$ vector of size 1. Note that ΨB is the $m \times 1$ vector, $1^T\Psi$B is the sum of all the elements of ΨB, or the total effect size of mixture exposure, and is the mean change in the outcome when the levels of all the exposures increase by one quartile simultaneously. A 95% confidence interval (*CI*) was estimated using the bootstrapping method.

Ordering Mechanism The order of exposure significantly influenced the estimation and extrapolation of $\beta(x)$. In this study, we propose three common ordering mechanisms, acknowledging that other sorting approaches may be selected based on specific research contexts.

I. Customized ordering: This method allows researchers to order exposures based on their understanding of the research background. It leverages domain expertise to create meaningful sequences of exposures.

II. Correlation-based ordering: This approach orders exposures based on their interrelationships, positioning highly correlated variables close to one another. For example, hierarchical clustering can be employed to arrange variables according to the resulting dendrogram. Ordering results can vary significantly depending on the definition of the distance matrix.

i. Define $1 - \Sigma$, where Σ is the pairwise correlation matrix of exposures, as the distance matrix; then, exposures with high positive correlations are positioned closer, whereas those with strong

negative correlations are positioned farther apart.

ii. Define $1-|\Sigma|$ as the distance matrix; then, exposures with strong correlations (either positive or negative) are placed closer together, whereas those with correlations closer to 0 are positioned farther apart.

III. Association-based ordering: This method orders exposures based on the strength of their associations with an outcome. Variables with similar levels of association with the outcomes were positioned closer to each other.

Unlike traditional regression methods, where multicollinearity can lead to unstable parameter estimates, the GFLM addresses this challenge through its functional smoothing approach. When highly correlated exposures are ordered adjacently, their effects are smoothed together through a basis function expansion, effectively treating them as functional units rather than competing independent variables. This smoothing naturally regularizes the effect estimates, preventing the instability typically observed in standard regression approaches.

Hypothesis Test Testing the mixture effect of exposures in Model (2) is equivalent to testing hypothesis $H_0: \beta(x) = 0$ versus $H_1: \beta(x) \neq 0$. If $\beta(x) = 0$, the outcome is unrelated to the mixture of exposures; otherwise, an association exists. By expanding $\beta(x)$ under OLS into Model (4), the test hypothesis is converted to $H_0: \mathrm{B} = 0$ versus $H_1: \mathrm{B} \neq 0$. This transformation shifts the hypothesis by determining whether a continuous function equals zero when testing finite-dimensional parameters. Based on the testing principle of nested models, we can construct likelihood ratio statistics (LRTs) as

$$\lambda\left(Q^{'}\right) = \frac{\sup_{\beta(x)=0} L\left(\beta(x)\,|Q^{'}\right)}{\sup_{\beta(x)} L\left(\beta(x)\,|Q^{'}\right)}, \tag{6}$$

and $-2\ln\left(\lambda\left(Q^{'}\right)\right)$ follows an χ^2 distribution with degrees of freedom K.

The sequence kernel score test, also known as the global test (GT), is an alternative test statistic suitable for nested models. Within this framework, α in Model (4) is treated as a fixed effect, whereas $\mathrm{B} = (\beta_1, \cdots, \beta_K)^T$ is assumed to be an independent and identically distributed random effect, following a normal distribution with a mean of 0 and variance τ^2. The test hypothesis is subsequently transformed into $H_0: \tau^2 = 0$ versus $H_1: \tau^2 \neq 0$. If $\tau^2 = 0$, each element in B equals 0, indicating no association between the exposures and the outcome. Denote $W = (W_1, \cdots, W_n)^T$, $\mathcal{K} = WW^T$, and the variance-component functional kernel score test statistic is

$$S\left(\hat{\mu}, \hat{\sigma}_e^2\right) = \frac{(Y-\hat{\mu})^T \mathcal{K} (Y-\hat{\mu})}{\hat{\sigma}_e^2}, \tag{7}$$

where $\hat{\mu}$ and $\hat{\sigma}_e^2$ are the predicted mean and variance under the null hypothesis, respectively. That is, $\hat{\mu} = \hat{\alpha}_0 + Z^T\hat{\alpha}$, in which $Z = (Z_1, \cdots, Z_n)^T$ is the covariate matrix, $\hat{\alpha}_0$ and $\hat{\alpha}$ are the estimations under the null hypothesis; then, $S\left(\hat{\mu}, \hat{\sigma}_e^2\right)$ follows a χ^2 distribution, which can be approximated by an $\delta\chi_v^2$ distribution with a degree of freedom v and scale parameter δ[20-23]. Equation (6) is as follows:

$$E\left(S\left(\mu, \sigma_e^2\right)\right) = tr(\mathcal{K}),\ Var\left(S\left(\mu, \sigma_e^2\right)\right) = 2tr\left(\mathcal{K}^2\right). \tag{8}$$

in which μ and σ_e^2 are substituted by $\hat{\mu}$ and $\hat{\sigma}_e^2$, respectively. According to Kwee et al.[24], $\mathrm{E}\left(S\left(\mu, \sigma_e^2\right)\right) = tr(\mathcal{K})$ can be estimated by $\hat{e} = tr(P_0\mathcal{K})$, where $P_0 = I_n - Z\left(Z^TZ\right)^{-1}Z^T$ and where I_n is an $n \times n$ identity matrix. Furthermore, the variance $Var\left(S\left(\mu, \sigma_e^2\right)\right) = 2tr\left(\mathcal{K}^2\right)$ can be estimated by $\hat{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma^2}^2 / I_{\sigma^2\sigma^2}^2$, where $I_{\tau\tau} = 2tr\left((P_0\mathcal{K})^2\right)$, $I_{\tau\sigma^2}^2 = 2tr(P_0\mathcal{K}P_0)$, and $I_{\sigma^2\sigma^2}^2 = 2tr\left(P_0^2\right)$. Solving equations $\delta v = \hat{e}$ and $2\delta^2 v = \hat{I}_{\tau\tau}$ yields approximations of the scale parameter and the degree of freedom as follows:

$$\delta = \frac{\hat{I}_{\tau\tau}}{2\hat{e}},\ v = \frac{\hat{I}_{\tau\tau}}{2\delta^2} = \frac{2\hat{e}^2}{\hat{I}_{\tau\tau}}. \tag{9}$$

An R package implementing GFLM is available at https://github.com/Peng247/Research and can be installed using the R command *devtools::install_github ("Peng247/Research").*

Numerical Simulation

To evaluate the performance of the GFLM in mixture exposure studies, an extensive simulation study was conducted using the NHANES dataset as the data pool. This approach allows the maintenance of realistic correlation structures and distributions of exposure, thereby providing a more practical assessment of the capabilities of the method. To evaluate GFLM's performance of the GFLM relative to the existing methods, we included WQS regression in the unidirectional simulation scenarios for comparison, as both methods can be directly compared in terms of Type I error control, statistical power, and effect estimation accuracy.

The dataset comprises the 2011–2012,

2013–2014, and 2015–2016 dataset[10]. The dataset comprised 5,960 adult participants aged 20–60 years, for whom reliable dietary data and complete information on relevant covariates were available. Individuals who had experienced weight loss or other health-related diets at the time of the survey were excluded. The exposure mixture in the data pool consisted of 37 nutrients estimated from dietary intake data collected through two 24-hour dietary recall interviews. These nutrients include a wide range of dietary components such as macronutrients (e.g., carbohydrates, proteins, and various types of fats), vitamins (e.g., vitamins A, complex vitamins, vitamins C, D, E, and K), minerals (e.g., calcium, iron, magnesium, and sodium), and other dietary factors (e.g., fiber and caffeine). The outcome of interest in the original study was obesity, defined as a body mass index (BMI) greater than or equal to 30 kg/m^2. Approximately 36.2% of individuals in the dataset were classified as obese. The pairwise Spearman correlation coefficients of nutrients are shown in Figure 1, with correlation coefficients ranging from −0.08 to 0.883.

Type I Error Simulation To evaluate the robustness of GFLM, we estimated the empirical type I error rate using a simulation approach. Random samples of 1,000, 1,500, and 2,000 participants were drawn from the original dataset. Within each sample, the order of the outcome variable BMI (both continuous and dichotomous) was randomly permuted to

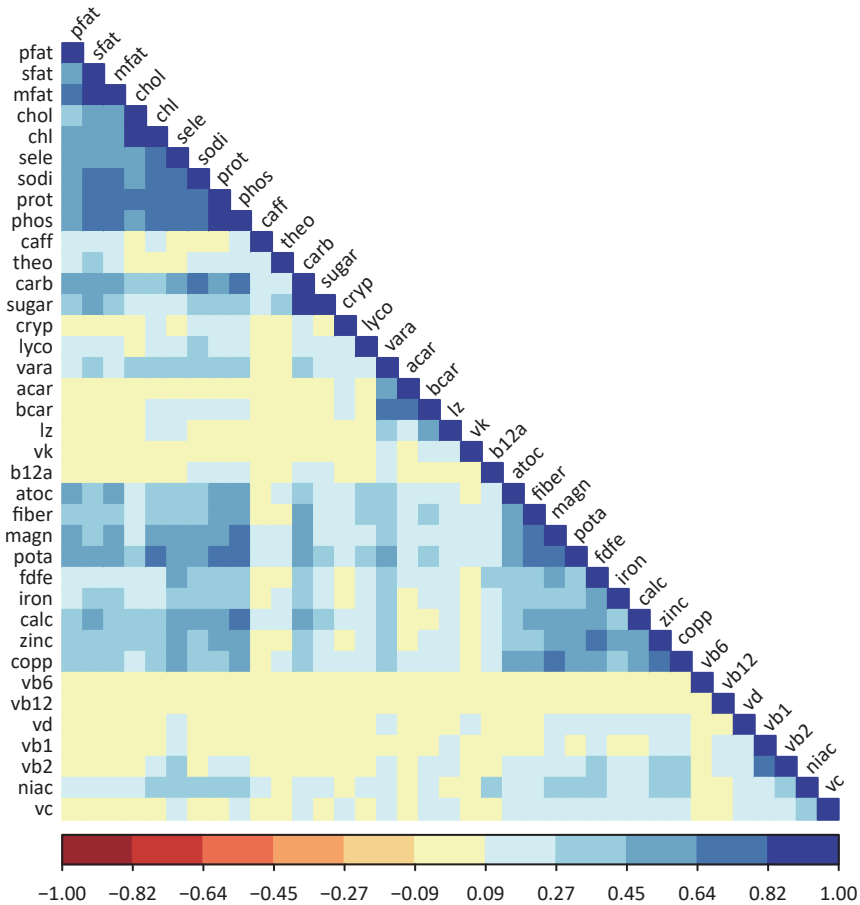

Figure 1. Spearman correlation matrix among nutrients. theo: Theobromine; caff: Caffeine; sele: Selenium; pota: Potassium; sodi: Sodium; copp: Copper; zinc: Zinc; iron: Iron; magn: Magnesium; phos: Phosphorus; calc: Calcium; vk: Vitamin K; vd: Vitamin D (D2+D3); vc: Vitamin C; b12a: vitamin B12; vb12: Vitamin B12; chl: Cholesterol; fdfe: Folate; zine: Zinc; vb6: Vitamin B6; niac: Niacin; vb2: Riboflavin (Vitamin B2); vb1: Thiamin (Vitamin B1); lz: Lutein + zeaxanthin; lyco: Lycopene; cryp: Beta-cryptoxanthin; bcar: Beta-carotene; acar: Alpha-carotene; vara: Vitamin A, RAE; atoc: Vitamin E as alpha-tocopherol; chol: Total choline; pfat: Total polyunsaturated fatty acids; mfat: Total monounsaturated fatty acids; sfat: Total saturated fatty acids; fiber: Dietary fiber; sugar: Total sugars; carb: Carbohydrate; prot: Protein.

disrupt potential associations with exposures. GFLM were constructed using age and sex as covariates and 37 nutrients as mixture exposures. The significance of the mixture exposure effect was tested using both the LRT and GT.

Given the random permutation of BMI, no significant association between mixture exposure and outcomes was expected. Therefore, the proportion of significant results across simulations provided an estimate of the type-I error rate. Denote P_k as the p value of the ith run of simulation and α' as the predetermined significance level; then, the empirical type I error rate after 10^4 runs is

$$\hat{\alpha} = \frac{\sum_{k=1}^{10^4} I\left(P_k < \alpha'\right)}{10^4}. \quad (10)$$

Power Simulation Under the alternative hypothesis, mixed exposure is associated with the outcome. The continuous outcome is generated by

$$E(y_i) = \alpha_0 + w_1 Z_{1i} + w_2 Z_{2i} + w_3 \sum_{i=1}^{m_+} v_i p_i - w_4 \sum_{j=1}^{m_-} u_j q_j + \epsilon_i, \quad (11)$$

in which $Z_{1i}\ N(0,1)$ and $Z_{2i}\ Bernoulli(0.5)$ are continuous and binary covariates, respectively; m_+ and m_- are the numbers of two sets of causal exposures p_i and q_j with weights v_i and u_j restricted to $\sum_{i=1}^{m_+} v_i = \sum_{j=1}^{m_-} u_j = 1$; $w_k \in [0,1)$ with $\sum_{k=1}^{4} w_k = 1$ are the effect sizes of the corresponding terms; and $\epsilon_i\ N(0,1)$ are random errors. In Model (7), the negative sign of w_4 indicates that q_j represents exposure with negative effects on the outcome. Given the mutual exclusivity of p_i and q_j, this model assumes a unidirectional association between any single exposure and outcome. The proportion of null-effect exposures can be controlled by adjusting the magnitudes of m_+ and m_-. Therefore, w_3 and w_4 represent the cumulative positive and negative exposure effects, respectively, and their difference $w_3 - w_4$ denoting the overall mixture exposure effect. By setting $\alpha_0 = \log(0.25)$, corresponding to a 20% incidence rate, the continuous outcome is generated by $y_i\ N(E(y_i),1)$, and the binary outcome is generated by $y_i\ Bernoulli\left(\exp\{E(y_i)\} / (\exp\{E(y_i)\} + 1)\right)$.

In the empirical power simulation, we focused on two scenarios of interest: unidirectional ($w_4 = 0$) and bidirectional ($w_4 \neq 0$) exposure effects. The effect size was defined as the proportion of the total outcome variance attributable to the mixture exposure effect; 1/3, 1/4, and 1/6 were selected for the simulation. The causal proportion, which represents the ratio of causally active exposures to the total number of exposures in the mixture, was used to model the different patterns of exposure effects. This proportion ranges from 1 (indicating that all exposures are associated with the outcome) to 1/16 (intermediate values of 1/2, 1/4, and 1/8). Sample sizes of 1,000, 1,500, and 2,000 were selected, with 10^3 simulation runs for each empirical power calculation. Following the same notation as the empirical type-I error rate, the empirical power is

$$1 - \hat{\beta} = \frac{\sum_{k=1}^{10^3} I\left(P_k < \alpha'\right)}{10^3}. \quad (12)$$

Real Data Analysis To demonstrate the practicability of the GFLM, we applied two datasets from the NHANES, each representing a different type of mixture exposure scenario.

The first was the dataset mentioned in the numerical simulation section. The primary objective was to investigate the impact of 37 nutrients on obesity, while controlling for age, sex, race, exercise intensity, and smoking status. Using the GFLM, we tested the significance of the nutrient mixture effect, estimated the magnitude of the overall mixture effect, and quantified the individual contributions of each nutrient.

The second dataset was derived from six NHANES cycles spanning 2007–2018 whose objective was to investigate the association between per- and polyfluoroalkyl substance (PFAS) exposure and gout risk. The initial sample comprised 59,842 participants, which were refined to 7,101 participants through a series of exclusion criteria. The primary outcome (gout status) was determined on the basis of self-reported data collected using a structured questionnaire. The exposure assessment focused on the serum concentrations of four PFAS: perfluorooctanoic acid (PFOA), perfluorooctane sulfonic acid (PFOS), perfluorohexane sulfonic acid (PFHxS), and perfluorononanoic acid (PFNA). These compounds were quantified using online solid-phase extraction coupled with high-performance liquid chromatography-turbo ion spray tandem mass spectrometry. A comprehensive set of covariates was included to adjust for the potential confounding effects. These included demographic variables (age, sex, and ethnicity), socioeconomic factors (educational attainment and family poverty-income ratio), lifestyle factors (smoking status, alcohol

consumption, and physical activity level), and health-related variables (BMI). We aimed to test the individual and mixed effects of PFAS on the risk of gout while controlling for various covariates. The detailed sample selection process, definition of outcomes, exposures, covariates, and statistical description of covariates by gout status are presented in Supplementary Information.

RESULTS

Type I Error

The results of the type-I error simulations are listed in Table 1. The dimension reduction rate indicates the intensity of the dimensionality reduction, with 1/2 indicating a reduction of half the original dimensions. In Model (3), the EEF $\beta(x)$ is expanded by finite-dimensional basis functions and coefficients ΨB, with the approximation controlled by the number of basis functions K. The dimension reduction rate was calculated as K/m. Smaller K values represent greater dimensionality reduction, resulting in poorer fitting compared to larger K values. Table 1 illustrates the robustness of the GFLM in analyzing continuous and binary outcomes under various sample sizes, significance levels, dimensionality reduction intensities, and ordering mechanisms.

Both LRT and GT generally maintain type I error rates close to the nominal significance level α, with GT exhibiting a more conservative tendency. Under association-based ordering, both tests maintain stable control with rates ranging from 0.049–0.059 at $\alpha = 0.05$. Correlation-based ordering showed slightly more variability in the control group, particularly for the GT with a smaller sample size. The GT was insensitive to the degree of dimensionality reduction, whereas the LRT tended toward an inflated type I error as the number of mixture exposures increased. For LRT, the robustness of testing continuous outcomes appeared slightly superior to that of binary outcomes, whereas GT demonstrated comparable control across both outcome types. In summary, these findings suggest that the GT provides a more robust testing method for GFLM, maintaining better type I error control across various scenarios of dimensionality outcome types and ordering mechanisms.

Power

The empirical power of the GFLM was evaluated through extensive simulations, and the results are shown in Figures 2–5. These simulations compared the performances of the LRT and GT under various scenarios, including unidirectional and bidirectional exposure effects for both continuous and binary

Table 1. Simulation results of type I error rates of LRT and GT of the GFLM models for continuous and binary outcomes

Ordering mechanism	Sample size	α level	Dimension reduction rate = 1/4				Dimension reduction rate = 1/2			
			LRT (CTN)	GT (CTN)	LRT (BNY)	GT (BNY)	LRT (CTN)	GT (CTN)	LRT (BNY)	GT (BNY)
Association based ordering	1,000	0.05	0.055	0.051	0.058	0.051	0.056	0.050	0.059	0.050
		0.01	0.013	0.010	0.012	0.010	0.013	0.011	0.013	0.010
	1,500	0.05	0.052	0.050	0.059	0.051	0.055	0.051	0.058	0.050
		0.01	0.012	0.010	0.013	0.012	0.013	0.011	0.013	0.010
	2,000	0.05	0.054	0.051	0.055	0.051	0.056	0.049	0.057	0.050
		0.01	0.012	0.011	0.011	0.010	0.013	0.009	0.013	0.010
Correlation based ordering	1,000	0.05	0.052	0.044	0.049	0.051	0.042	0.041	0.055	0.031
		0.01	0.012	0.011	0.014	0.011	0.010	0.008	0.008	0.002
	1,500	0.05	0.053	0.055	0.052	0.049	0.048	0.054	0.053	0.050
		0.01	0.008	0.011	0.011	0.009	0.011	0.012	0.011	0.010
	2,000	0.05	0.053	0.042	0.053	0.051	0.053	0.055	0.055	0.047
		0.01	0.013	0.008	0.011	0.012	0.008	0.009	0.008	0.008

Note. LRT, likelihood ratio test; GT, global test; CTN, continuous; BNY, binary. GFLM, generalized functional linear model.

outcomes.

In the unidirectional exposure scenarios (Figures 2–3), the GFLM demonstrated greater power for continuous outcomes than for binary outcomes. The GT exhibited superior overall performance relative to the LRT, particularly in terms of robustness to effect size variations. This suggests that the GT may be preferable in situations where the magnitude of the exposure effects is uncertain under the unidirectional assumption of mixture exposure. A comparison with WQS regression reveals that the GFLM provides comparable or superior power across different scenarios, particularly for binary outcomes and smaller effect sizes. While both methods maintain good power for continuous outcomes, the GFLM demonstrates a more stable performance across varying causal proportions. Bidirectional exposure simulations (Figures 4–5) revealed that the LRT outperformed the GT. This performance difference was particularly pronounced for binary outcomes, particularly when the association-based ordering of exposures was employed. These findings highlight the importance of selecting an appropriate test statistic based on the anticipated direction of the exposure effects and the nature of the outcome variable.

In contrast to the intuitive expectation that opposing effects in a mixture would lead to reduced detectability, the simulations demonstrated that the absolute sum of the effect sizes is a key determinant of statistical power. Figure 5 illustrates that regardless of the proportion of positive to negative effects, larger absolute effect sizes consistently yielded greater power. This result underscores the ability of the GFLM to detect significant mixture effects, even in the presence of opposing individual exposure effects.

Effect Size

Figures 6 and 7 present the results of effect size estimation for various simulation scenarios. The red dashed line represents the true effect size, and the solid lines represent the 95% *CI*s for each simulation setting. Across all the simulated scenarios, the 95%

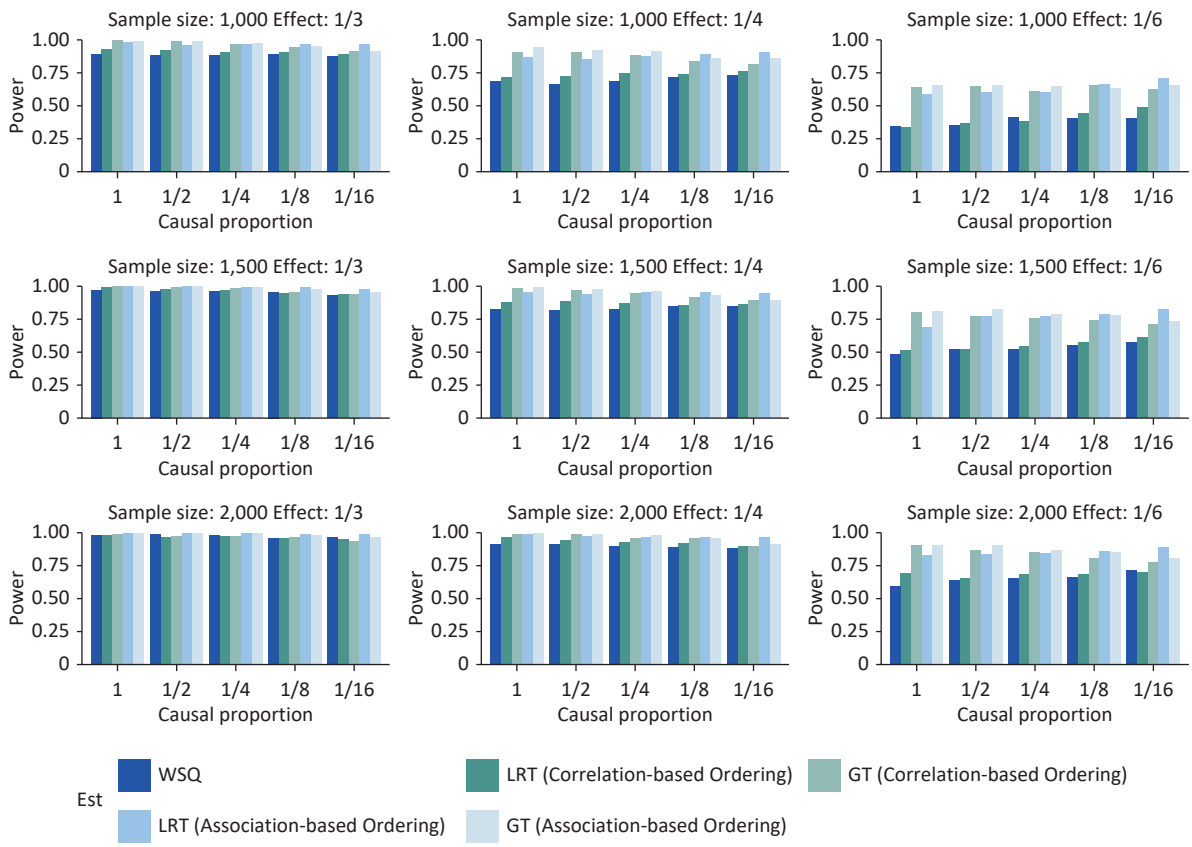

Figure 2. Simulation of the unidirectional exposure effect for continuous outcomes under various sample sizes, total effect sizes, causal exposure proportions, and testing method combinations. WQS, weighted quantile sum regression; LRT, likelihood ratio test; GT, global test.

*CI*s consistently encompassed the true effect size, with point estimates generally clustered around the true value. As expected, increasing the sample size reduced the width of the *CI*s. Figure 6 shows that the GFLM provides more reliable estimates than the WQS, which tends to overestimate mixture effects, particularly for binary outcomes. GFLM's confidence intervals of the GFLM consistently contain the true effect size across different scenarios, demonstrating its robust performance in effect estimation. Notably, Figure 6 shows that smaller true mixture effects correspond to narrower *CI*s. This inverse relationship suggests that the GFLM performs particularly well in detecting and estimating subtle mixture effects.

In the simulation studies, we observed that for continuous outcomes, while the 95% *CI*s of the mixture effect size estimates consistently encompassed the true values, the point estimates deviated from them. Further investigation revealed that a two-step approach could significantly increase the accuracy of the point estimates. This approach involves normalizing the continuous outcome, converting it into a binary variable, and then refitting the model. The final estimate was obtained by averaging the total effect estimates from both original and transformed models. Figures 6 and 7 present the results of the mixed-averaging method. For binary outcomes, the point estimates demonstrated remarkable accuracy across various simulation scenarios, closely aligning with the true values. However, it is worth noting that the 95% *CI*s for binary outcomes were generally wider than those observed for continuous outcomes.

Data Analysis

Nutrients-BMI Association The GFLM model was applied to reanalyze data from the NHANES 2011–2016 cycles to examine and estimate the effects of 37 nutrients on BMI, while also demonstrating the relative contribution of each nutrient. Analysis of variance inflation factors (VIF) revealed correlation among the nutrients, with 9 nutrients showing high correlation (VIF > 5) and 14 showing moderate correlation (2 < VIF ≤ 5). Table 2 presents the estimated mixture effects, 95% *CI*s, and test results for the different exposure ordering

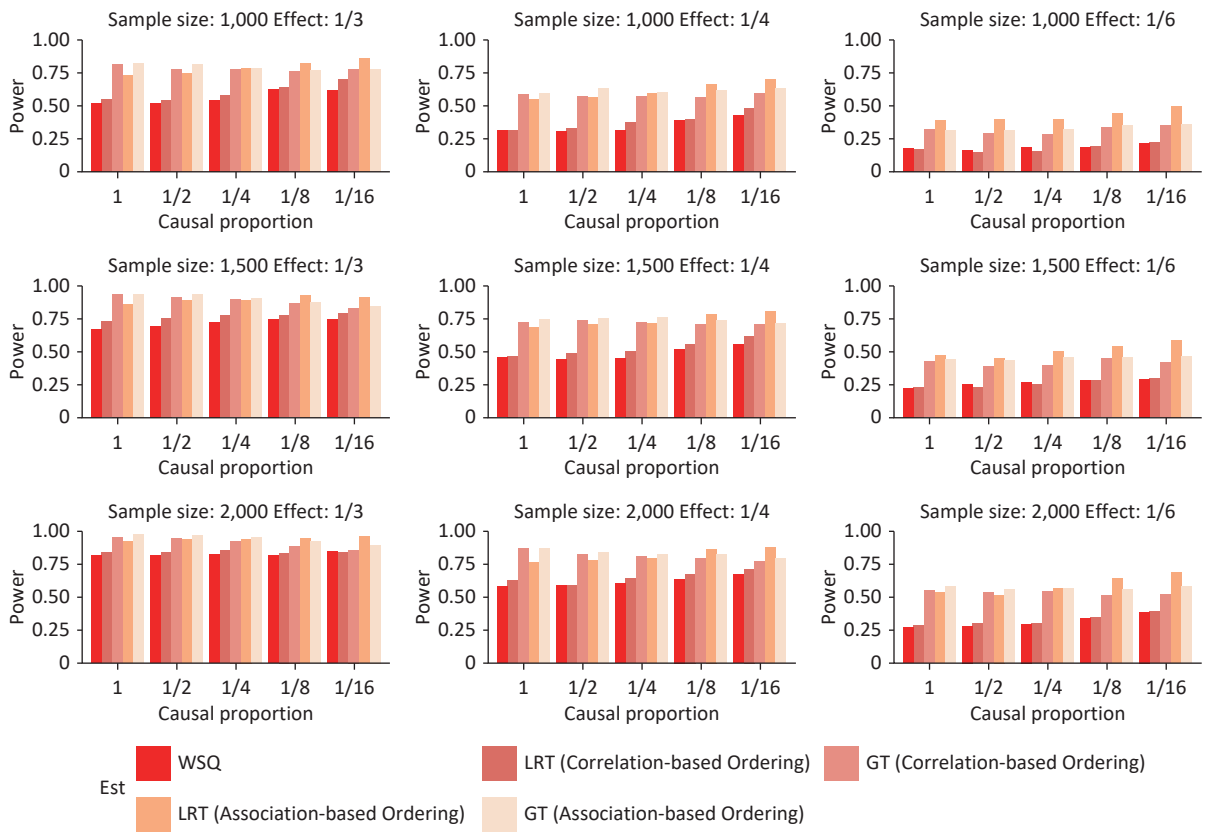

Figure 3. Power simulation of the unidirectional exposure effect for binary outcomes under various sample sizes, total effect size magnitudes, causal exposure proportions, and testing method combinations. WQS, weighted quantile sum regression; LRT, likelihood ratio test; GT, global test.

mechanisms. After controlling for covariates, the nutrient mixture had a significant effect on BMI. Under association-based ordering, a one-quartile increase across all nutrients corresponded to a 0.246 unit decrease in BMI. Under correlation-based ordering, this effect was estimated as a 0.323-unit decrease.

Figure 8 illustrates the effect function calculated using model (3) for both ordering mechanisms. While Renzeitt's[10] work identified magnesium and sodium as the nutrients with the largest negative and positive contributions, respectively, the GFLM analysis suggests a more nuanced interpretation. Although these nutrients have both negative and positive effects, they are not significant contributors. Instead, their effects appear to be a part of a complex interplay with other nutrients. Figure 8 shows that fiber and fat consistently emerged as nutrients with the greatest negative and positive effects, respectively, regardless of the ordering method. This finding is consistent with those of numerous previous studies on the effects of diet on BMI[25-27].

Notably, for certain nutrients, the estimated direction of the effect differed between the two ordering mechanisms. This discrepancy is related to the GFLM estimation methodology and warrants further discussion in the subsequent sections of this paper.

PFAS-gout Association The GFLM model was applied to analyze data from the NHANES 2007–2018 cycles to investigate the association between the four PFASs and gout risk while controlling for various covariates.

Analysis of the dataset revealed strong linear correlations between the four PFAS, as illustrated by the Pearson correlation coefficient matrix in Table 3. A multistep analytical approach was used to examine the relationship between PFAS exposure and gout risk. Initially, logistic regression models were used to assess the association between individual PFAS and gout after adjusting for all covariates. Subsequently, a multivariate logistic model incorporating all the four PFAS was constructed to explore their combined effects. The results of these analyses are presented in the "individual exposure analysis" section in Table 4. Both the univariate and multivariate analyses showed no significant

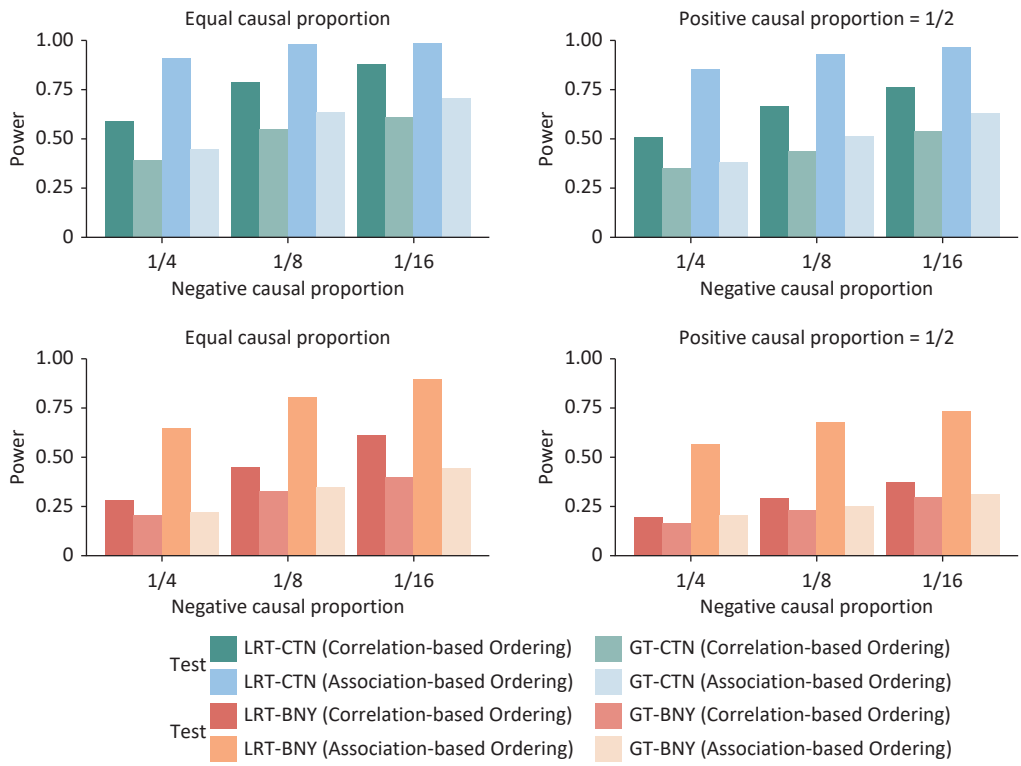

Figure 4. Power simulation of the bidirectional exposure effect for continuous (first row) and binary (second row) outcomes under various causal proportion combinations when the sample size is fixed to 1,500 and when the positive effect size is assumed to be equal to the negative effect size. WQS, weighted quantile sum regression; LRT, likelihood ratio test; GT, global test.

association, with only PFOA exhibiting marginal significance in the multivariate model. However, recognizing the limitations of logistic regression in capturing the mixture effects, further analysis was conducted using both WQS and the proposed GFLM. The results are presented in the lower half of Table 4.

WQS regression analysis suggested a significant association between the joint effect of the four PFAS and gout risk (*OR*, 1.40; 95% *CI*: 1.31–1.50). In contrast, the GFLM showed no significant association. To interpret this discrepancy in the results, we noticed high positive correlations among the PFAS, as shown in Table 3, which suggests that multicollinearity may significantly influence parameter estimation in the WQS framework. Moreover, the discrepancy can be understood through simulation studies, which demonstrate that WQS tends to overestimate the mixture effects for binary outcomes. Given this tendency and the fact that GFLM's demonstrated robust performance in the simulations, the GFLM results suggesting no significant PFAS mixture effect on gout risk may be more reliable.

DISCUSSION

In this study, we propose a novel model framework for analyzing mixture exposure data. Drawing inspiration from the FDA, we introduce the concept of fitting the exposure effects as continuous functions. This approach allows for efficient and accurate effect estimation, while fully capturing the correlations between exposures. Statistical simulations demonstrated that the proposed GFLM model exhibited a robust performance and preferable statistical power across various settings. The effect size estimates were reliable, with 95% *CI*s consistently encompassing the true values across all simulation scenarios. We applied the model to reanalyze the effects of 37 nutrients on BMI using the NHANES database, which yielded results that

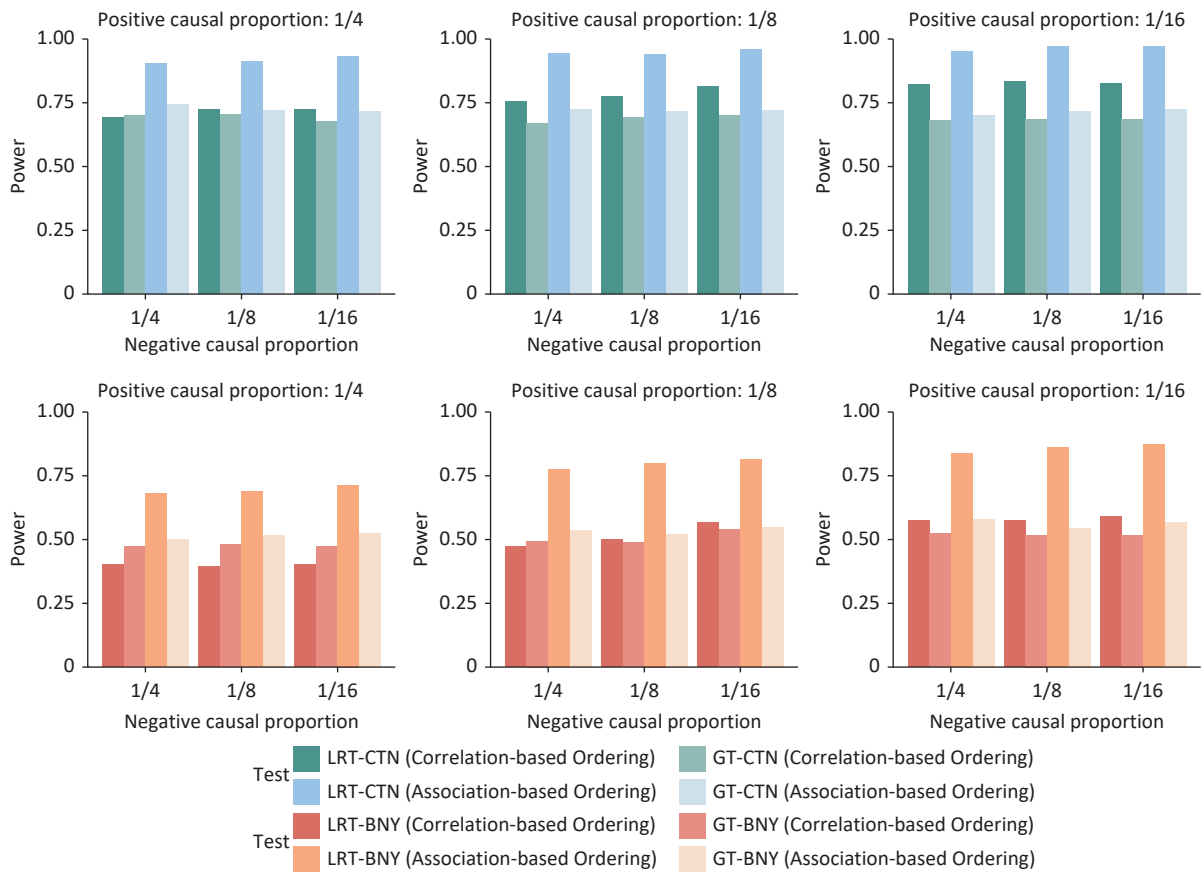

Figure 5. Power simulation of the bidirectional exposure effect for continuous (the first row) and binary (the second row) outcomes under various causal proportion combinations when the sample size is fixed to 1,500 and assuming that the positive effect size is twice the negative effect size. WQS, weighted quantile sum regression; LRT, likelihood ratio test; GT, global test.

were more interpretable than those of the original study. Additionally, the application of the GFLM to the PFAS-gout dataset further demonstrated its utility in handling complex, correlated exposures. Unlike the WQS regression, which suggested a significant association between the PFAS mixture and gout risk, the GFLM showed no significant association. This discrepancy highlights GFLM's robustness of the GFLM to multicollinearity, which is a common challenge in environmental mixture analysis.

Given the innovative integration of FDA techniques into mixture exposure analyses, it is essential to highlight the key technical considerations for practical applications. Typically, functional regression is applied when variables are functions of time or other continua; however, exposure generally lacks inherent order. To address this, our analytical framework first orders the exposures according to specific rules and then approximates the position as a continuous quantity for function fitting. The scientific basis for this approach stems from observations in time-series data, where adjacent points often exhibit the highest correlation, leading to our proposed correlation-based ordering mechanism, which arranges exposures based on the hierarchical clustering of their associations. Although the selection of basis functions is an important consideration in the FDA, our simulation results demonstrate that the GFLM maintains a robust performance across different choices of basis function numbers. We recommend starting with $K = m/4$ basis functions for general applications, where m is the number of exposures, with adjustments based on cross-validation results. Importantly, the validity of GFLM's conclusions of the GFLM remained stable across different dimensionality reduction rates.

The implications of the two proposed ordering methods are notable. Although correlation-based ordering satisfies the conditions for functional regression, studies have shown that two positively associated exposures or biomarkers can have opposing effects on the outcome[28,29]. This phenomenon explains why some exposures in our data analysis exhibited opposite effects under

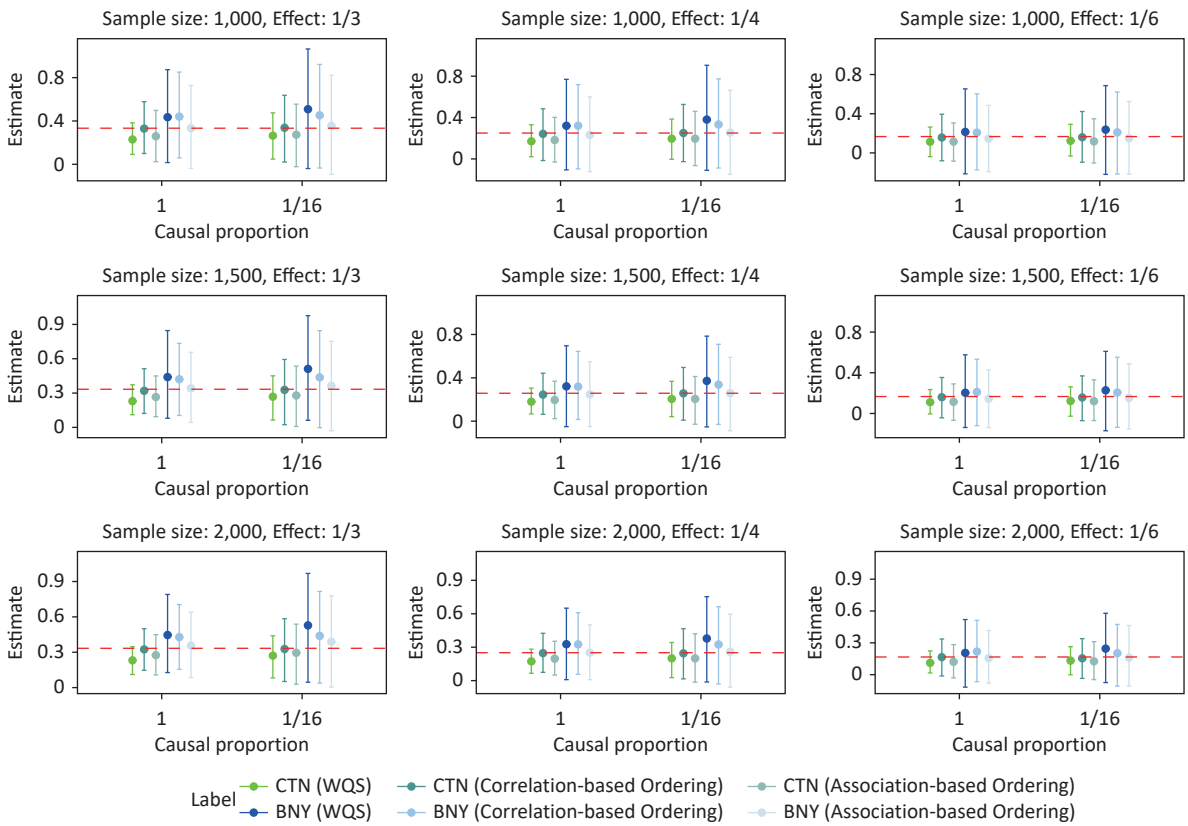

Figure 6. Mixture exposure effect estimates with 95% confidence interval of unidirectional effects for continuous and binary outcomes under various sample sizes, causal effect sizes, and causal proportion setting combinations. CTN, continuous; BNY, binary; WQS, weighted quantile sum regression.

different ordering mechanisms. Therefore, we also propose an ordering based on the strength of the association with the outcomes. In practical applications, researchers should carefully consider the potential mechanisms of the exposure effects and select the most appropriate ordering method.

The choice between LRT and GT should be guided by study characteristics and research priorities. While the LRT shows slightly elevated Type I error rates but higher power when the effects are concentrated in fewer exposures, the GT maintains strict control and performs better when the effects are distributed across multiple exposures. GT is recommended when strict type-I error control is required, whereas LRT may be preferable for exploratory analyses or when higher sensitivity is acceptable.

Our study findings provide guidance for interpreting effect size estimates. In real-world scenarios, mixture exposure often includes near-zero-effect exposure. However, the functional regression fitting process may assign small pseudo-effects to these exposures to maintain the functional smoothness. The strength of the GFLM lies in identifying consistent patterns and peak effects across different ordering mechanisms rather than precisely estimating individual effects. Traditional regression approaches are appropriate for investigating specific exposure effects. When using the GFLM, we recommend focusing on effects that remain consistent across different ordering mechanisms, as these provide stronger evidence of true associations. Visualizations such as those in Figure 8 can be used to observe the relative

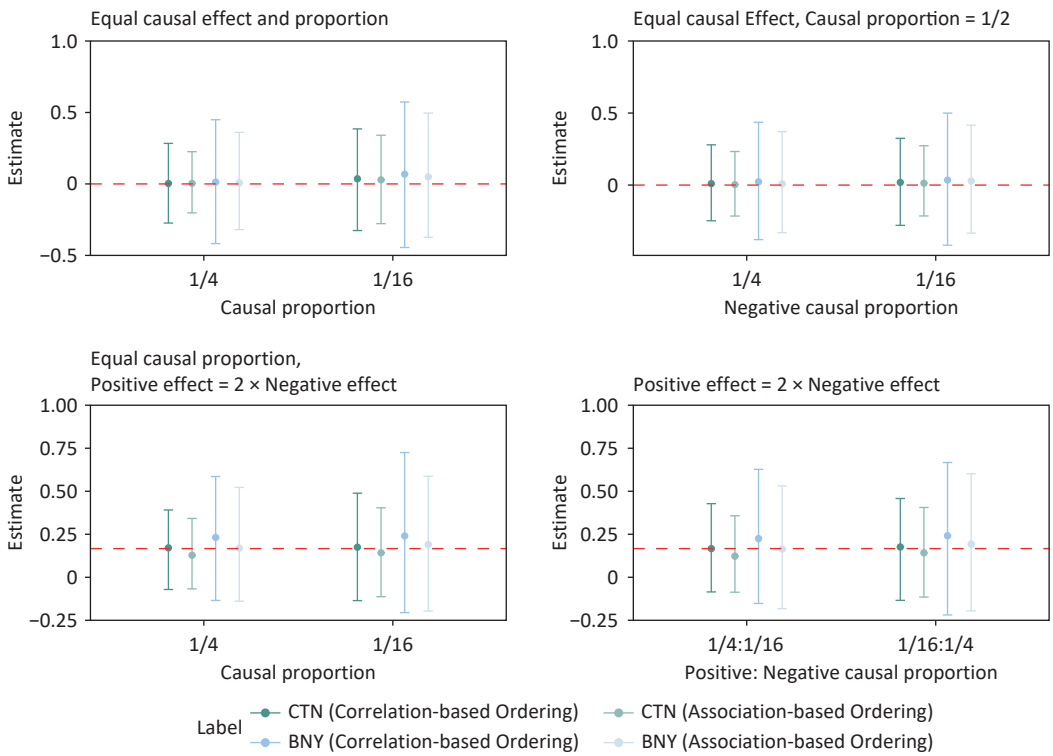

Figure 7. Mixture exposure effect estimate with 95% confidence interval of the bidirectional effect for continuous and binary outcomes under various sample sizes, causal effect sizes, and causal proportion setting combinations. CTN, continuous; BNY, binary; WQS, weighted quantile sum regression.

Table 2. Analysis results of the effects of 37 nutrient mixtures on BMI (NHANES 2011–2016)

Ordering mechanism	Estimate	95% *CI*	P_{LRT}	P_{GT}
Association-based	−0.246	−0.5000, −0.012	6.6×10^{-5}	5.3×10^{-3}
Correlation-based	−0.323	−0.6200, −0.0079	3.3×10^{-4}	0.012

Note. *CI*, confidence interval; LRT, likelihood ratio test; GT, global test.

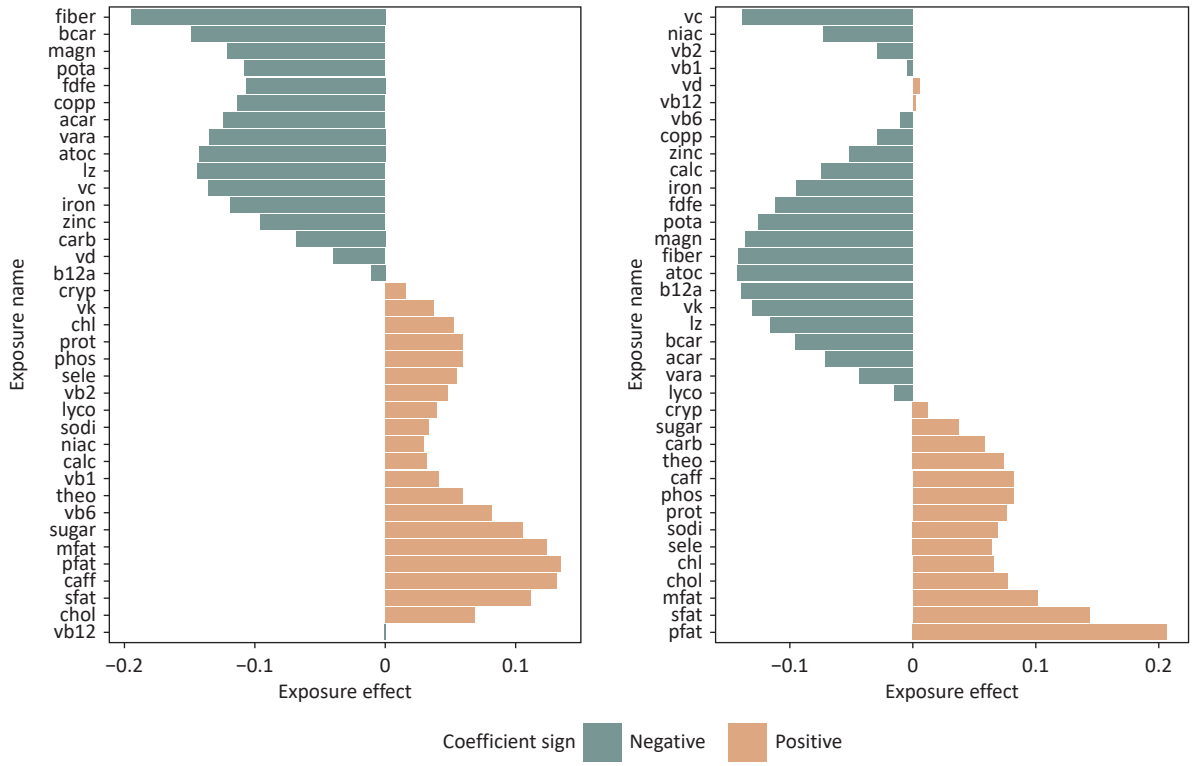

Figure 8. Nutrient effect estimates under association-based and correlation-based ordering assumptions (NHANES 2011–2016).

Table 3. Pearson correlation between four PFAS levels after $\log_2$ transformation

Correlation	PFNA	PFOA	PFOS
PFOA	0.72		
PFOS	0.73	0.69	
PFHx S	0.46	0.62	0.67

Note. PFOA, perfluorooctanoic acid; PFOS, perfluorooctane sulfonic acid; PFHxS, perfluorohexane sulfonic acid; PFNA, perfluorononanoic acid.

Table 4. Mixture effect estimates of logistic regression, WQS, and GFLM

Analysis	Model	Mixture effect (95% *CI*)			
	Logistic Regression	PFOA	PFOS	PFHxS	PFNA
Individual exposure analysis	Univariate	1.03 (0.91, 1.16)	0.92 (0.84, 1.02)	0.92 (0.83, 1.02)	0.98 (0.87, 1.09)
	Multivariate	1.24 (1.02, 1.49)	0.88 (0.75, 1.04)	0.89 (0.77, 1.03)	1.01 (0.83, 1.21)
	WQS	1.40 (1.31, 1.50)			
Mixture exposure analysis	GFLM (correlation-based ordering)	1.05 (0.79, 1.55)			
	GFLM (association-based ordering)	1.07 (0.77, 1.62)			

Note. *CI*, confidence interval; WQS, weighted quantile sum; GFLM, generalized functional linear model; PFOA, perfluorooctanoic acid;PFOS, perfluorooctane sulfonic acid; PFHxS, perfluorohexane sulfonic acid; PFNA, perfluorononanoic acid.

contributions of each exposure within the mixture.

The proposed GFLM method has several advantages. It provides results that align more closely with the researchers' expectations from statistical analyses. By leveraging the nested model framework and functional regression techniques, we calculated exact *P* values, mixture effect estimates, and *CI*s. Unlike the complex results of methods such as BKMR, these outputs are both familiar and easily interpretable. Second, the GFLM model exhibited superior computational efficiency. It is significantly faster than WQS, with computation times approximately 1/50 of the latter for sample sizes of 1,000 under the simulation settings described in Section 2.2. This advantage became even more pronounced with larger sample sizes. A detailed comparison of the running times is provided in the Supplementary file.

However, GFLM has some limitations. First, its theoretical foundation requires a higher level of statistical expertise from users, including programming proficiency. Second, the functional regression approach in Model (3) lacks standardized guidelines for selecting the basis function system, number, and order, thus requiring researchers to explore these aspects independently. In our simulations and data analysis, we used a quarter of the number of exposures and employed B-spline fitting for cubic functions, although Fourier basis functions might be more appropriate in some cases, depending on the specific research needs. Finally, for exposures with weak individual and mixture effects, their estimated contributions can be sensitive to the choice of ordering mechanism, necessitating careful interpretation based on the researchers' understanding of the problem.

In conclusion, the GFLM model presented in this study offers a theoretically sound and highly applicable method for analyzing mixture exposure. Statistical simulations and case analyses demonstrated their robustness, efficiency, and versatility. We have included R codes for the simulation and data analysis in the Appendix with the hope that this method will find widespread application in mixture exposure research, thereby advancing the field of public health.

Funding This research was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China (Grant Nos. 82304253) (and 82273709), the Foundation for Young Talents in Higher Education of Guangdong Province (Grant No. 2022KQNCX021), and the PhD Starting Project of Guangdong Medical University (Grant No. GDMUB2022054).

Competing Interests All authors declare no competing interests.

Ethics The original NHANES protocol was approved by the National Center for Health Statistics Research Ethics Review Board (Protocol #98-12, #2005-06, #2011-17, #2018-01). This study was conducted in accordance with the principles of the Declaration of Helsinki. All participants in the original NHANES provided written informed consent. This study, which used de-identified publicly available data, was exempt from institutional review board approval and did not require additional consent.

Authors' Contributions Conceptualization of the method, model proposal, design of simulation studies, and writing of the main manuscript text: Bingsong Zhang; Conducting simulation studies and collecting all results: Haibin Yu, Xin Peng, and Haiyi Yan; Data analysis and result collection: Siran Li, Shutong Luo, and Renhuizi Wei; Data collection and cleaning from NHANES: Zhujiang Zhou, Yalin Kuang, Yihuan Zheng, and Chulan Ou; Study supervision and thorough manuscript revision: Linhua Liu, Yuehua Hu, and Jindong Ni.

Acknowledgements We thank Dr. Ruzong Fan from Georgetown University, Dr. Chi-yang Chiu from the University of Tennessee Health Science Center, and Dr. Shuqi Wang from the University of Wisconsin, Madison for their kind advice regarding the methodology of this manuscript.

Consent for Publication All authors of this manuscript agree to its submission to Biomedical and Environmental Science for publication.

Data Sharing All the data are publicly available. Type I error simulation data and codes are provided in the supplementary file typr_I_error_simulation.R. The power simulation data and codes are provided in the supplementary file power_simulation.R. The generalized functional linear model is shown in the Supplementary file GFLM.R. The NHANES data can be found on its official website: https://www.cdc.gov/nchs/nhanes/index.htm. Data analysis of the PFAS-gout example is provided in the supplementary file. The supplementary materials will be available in www.besjournal.com.

Received: September 21, 2024;
Accepted: February 10, 2025

REFERENCES

1. Bopp SK, Barouki R, Brack W, et al. Current EU research activities on combined exposure to multiple chemicals.

Environ Int, 2018; 120, 544−62.
2. Yu LL, Liu W, Wang X, et al. A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture. Environ Pollut, 2022; 306, 119356.
3. Tanner EM, Bornehag CG, Gennings C. Repeated holdout validation for weighted quantile sum regression. MethodsX, 2019; 6, 2855−60.
4. Traini E, Huss A, Portengen L, et al. A multipollutant approach to estimating causal effects of air pollution mixtures on overall mortality in a large, prospective cohort. Epidemiology, 2022; 33, 514−22.
5. Carlin DJ, Rider CV, Woychik R, et al. Unraveling the health effects of environmental mixtures: an NIEHS priority. Environ Health Perspect, 2013; 121, A6−8.
6. Taylor KW, Joubert BR, Braun JM, et al. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environ Health Perspect, 2016; 124, A227−9.
7. Gibson EA, Nunez Y, Abuawad A, et al. An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length. Environ Health, 2019; 18, 76.
8. Chiu YH, Bellavia A, James-Todd T, et al. Evaluating effects of prenatal exposure to phthalate mixtures on birth weight: a comparison of three statistical approaches. Environ Int, 2018; 113, 231−9.
9. Carrico C, Gennings C, Wheeler DC, et al. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. J Agric Biol Environ Stat, 2015; 20, 100−20.
10. Renzetti S, Gennings C, Calza S. A weighted quantile sum regression with penalized weights and two indices. Front Public Health, 2023; 11, 1151821.
11. Keil AP, Buckley JP, O'Brien KM, et al. A quantile-based g-computation approach to addressing the effects of exposure mixtures. Environ Health Perspect, 2020; 128, 47004.
12. Bobb JF, Valeri L, Henn BC, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics, 2015; 16, 493−508.
13. Fan RZ, Wang YF, Mills JL, et al. Functional linear models for association analysis of quantitative traits. Genet Epidemiol, 2013; 37, 726−42.
14. Wu MC, Lee S, Cai TX, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet, 2011; 89, 82−93.
15. Ramsay JO, Silverman BW. Functional data analysis. Springer. 1997.
16. Ramsay J, Hooker G, Graves S. Functional data analysis with R and MATLAB. Springer. 2009.
17. Ferraty F, Romain Y. The Oxford handbook of functional data analysis. Oxford University Press. 2010.
18. de Boor C. A practical guide to splines, revised edition. Springer. 2001.
19. Horváth L, Kokoszka P. Inference for functional data with applications. Springer. 2012.
20. Duchesne P, De Micheaux PL. Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. Comput Stat Data An, 2010; 54, 858−62.
21. Davies RB. The distribution of a linear combination of χ^2 random variables. Appl Stat, 1980; 29, 323−33.
22. Liu H, Tang YQ, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data An, 2009; 53, 853−56.
23. Lin XH. Variance component testing in generalised linear models with random effects. Biometrika, 1997; 84, 309−26.
24. Kwee LC, Liu DW, Lin XH, et al. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet, 2008; 82, 386−97.
25. Buscemi J, Pugach O, Springfield S, et al. Associations between fiber intake and Body Mass Index (BMI) among African-American women participating in a randomized weight loss and maintenance trial. Eat Behav, 2018; 29, 48−53.
26. Carrasquilla GD, Jakupović H, Kilpeläinen TO. Dietary fat and the genetic risk of type 2 diabetes. Curr Diab Rep, 2019; 19, 109.
27. D'Angelo S, Motti ML, Meccariello R. ω-3 and ω-6 polyunsaturated fatty acids, obesity and cancer. Nutrients, 2020; 12, 2751.
28. Friedman AN, Fadem SZ. Reassessment of albumin as a nutritional marker in kidney disease. J Am Soc Nephrol, 2010; 21, 223−30.
29. Di Angelantonio E, Sarwar N, Perry P, et al. Major lipids, apolipoproteins, and risk of vascular disease. JAMA, 2009; 302, 1993−2000.