UC San Diego
UC San Diego Previously Published Works

Title

Testing departure from additivity in Tukey's model using shrinkage: application to a longitudinal setting

Permalink

https://escholarship.org/uc/item/3z086106

Journal

Statistics in Medicine, 33(29)

ISSN

0277-6715

Authors

Ko, Yi-An
Mukherjee, Bhramar
Smith, Jennifer A
et al.

Publication Date

2014-12-20

DOI

10.1002/sim.6281

Peer reviewed

Testing Departure from Additivity in Tukey's Model using Shrinkage: Application to a Longitudinal Setting

Yi-An Ko[a,*], **Bhramar Mukherjee**[a], **Jennifer A. Smith**[b], **Sung Kyun Park**[b,c], **Sharon L.R. Kardia**[b], **Matthew A. Allison**[d], **Pantel S. Vokonas**[e,f,g], **Jinbo Chen**[h], and **Ana V. Diez-Roux**[i]

[a]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A

[b]Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109, U.S.A

[c]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI 48109, U.S.A

[d]Department of Family and Preventive Medicine, University of California, San Diego, CA 92093, U.S.A

[e]VA Normative Aging Study, Veterans Affairs Boston Healthcare System, Jamaica Plain, MA 02130, U.S.A

[f]Department of Epidemiology, Boston University, Boston, MA 02118, U.S.A

[g]Department of Medicine, Boston University, Boston, MA 02118, U.S.A

[h]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, U.S.A

[i]Drexel University School of Public Health, Philadelphia, PA 19104, U.S.A

Abstract

While there has been extensive research developing gene-environment interaction (GEI) methods in case-control studies, little attention has been given to sparse and efficient modeling of GEI in longitudinal studies. In a two-way table for GEI with rows and columns as categorical variables, a conventional saturated interaction model involves estimation of a specific parameter for each cell, with constraints ensuring identifiability. The estimates are unbiased but are potentially inefficient because the number of parameters to be estimated can grow quickly with increasing categories of row/column factors. On the other hand, Tukey's one degree of freedom (df) model for non-additivity treats the interaction term as a scaled product of row and column main effects. Due to the parsimonious form of interaction, the interaction estimate leads to enhanced efficiency and the corresponding test could lead to increased power. Unfortunately, Tukey's model gives biased estimates and low power if the model is misspecified. When screening multiple GEIs where each genetic and environmental marker may exhibit a distinct interaction pattern, a robust estimator for interaction is important for GEI detection. We propose a shrinkage estimator for interaction effects that combines estimates from both Tukey's and saturated interaction models and use the

[*]Correspondence to: Yi-An Ko, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A., yianko@umich.edu.

corresponding Wald test for testing interaction in a longitudinal setting. The proposed estimator is robust to misspecification of interaction structure. We illustrate the proposed methods using two longitudinal studies — the Normative Aging Study and the Multi-Ethnic Study of Atherosclerosis.

Keywords

adaptive shrinkage estimation; gene-environment interaction; longitudinal data; Tukey's one df test for non-additivity

1. Introduction

The presence of gene-environment interactions (GEI) implies that the effect of an environmental exposure (E) is enhanced or reduced for a sub-group with a certain genotype or vice versa. Investigation of GEI is essential to better understand the etiology and development of common, complex diseases. Many longitudinal environmental epidemiology studies have been collecting genetic data with the goal of identifying GEI. In these cohort studies, GEI is often investigated by focusing on an established association between an exposure biomarker (e.g., lead levels in blood or bone) and a quantitative trait (e.g., pulse pressure), and how this association is modified by a selected set of genetic markers. The set of genes (candidate genes) to be studied is often determined by the metabolic pathway related to the exposure instead of an agnostic search across the genome.

While there has been extensive literature on GEI regarding ways to enhance the efficiency of interaction test in case-control studies [1, 2, 3], statistical methods for GEI in longitudinal settings remain limited. Methods to study disease-gene association in longitudinal settings, however, have started to receive attention. For instance, Wang *et al.* [4] proposed to estimate and test for time-varying genetic effects using semiparametric models with penalized splines. Fan *et al.* [5] also used penalized spline models to estimate the mean function and genetic regression coefficients with extensions to linkage disequilibrium (LD) mapping. Nevertheless, very limited number of studies have focused on testing of gene-gene interactions (GGI) or GEI for complex traits in longitudinal settings. The multivariate adaptive splines presented by Zhang [6, 7] have been applied to analyze GEI in longitudinal cohort studies (e.g., Zhu *et al.* [8]). Xu [9] developed an empirical Bayes method to estimate GGI effects under the mixed model framework and compared it with several variable selection procedures. Malzahn *et al.* [10] developed a nonparametric test for investigation of GGI in repeated measures data using a rank procedure. Mukherjee *et al.* [11] proposed to explore the GEI structure with various parsimonious classical ANOVA models for non-additivity by taking the average of repeated measurements and forming cell means of a two-way GEI table. Along the same lines, Ko *et al.* [12] extended the classical ANOVA models under a mixed model framework and developed a resampling-based test for GEI that accounts for correlation within repeated measures.

Typically, an interaction model including cross-product terms of gene and environment under the mixed model framework is used for testing GGI and GEI in longitudinal studies [13]. In considering the estimation of GEI for longitudinal data where both the genetic factor (G) and E are categorical variables, this conventional modeling approach involves distinct

parameter estimation for each configuration of GEI (i.e., a saturated interaction form) with sum-to-zero type constraints to ensure identifiability. Estimation bias is minimized since the model does not impose any structural assumptions on the interaction term. However, the number of parameters and hence the corresponding degrees of freedom (df) for the interaction test can become substantially large as the number of categories of G and/or E increases. In addition, under a saturated interaction model only observations in a cell can contribute to the parameter estimation for that cell. This may result in reduced efficiency and loss of power for detecting interactions because of small cell sample size in human studies involving a gene with a modest minor allele frequency.

Tukey's one df model for non-additivity [14], originally proposed for data with no replication per cell, has been applied to the modeling of GGI in cohort studies. Maity *et al.* [15] used Tukey's form of interaction for repeated measures data to test main genetic associations in the presence of GGI. The interaction term in Tukey's model is treated as a scaled product of main effects, implying that the existence of interaction is conditional on the presence of main effects. When a GEI study is based on a two-stage strategy, namely, the candidate genes are selected based on marginal genetic associations [16, 17], it may be reasonable to adopt Tukey's interaction form for GEI. Chatterjee *et al.* [18] proposed that Tukey's model is also consistent with the notion that individual markers within a gene are associated with disease through a common biological mechanism. However, when candidate genes are chosen in relation to an exposure pathway, genes may not necessarily have main effects. Also, when the assumption of Tukey's interaction structure is violated (e.g., absence of genetic main effects), the estimate for the interaction effect using Tukey's model will be biased and the corresponding one-df test can result in extremely low power [11, 19].

When searching for GEI across multiple genetic markers, it is possible that GEIs exhibit distinct interaction patterns, departing from Tukey's model. Conducting multiple tests under a fixed interaction structure (e.g., Tukey) may not capture interactions of alternative forms. At the same time, it would be advantageous to leverage the power of Tukey's test if it is indeed a plausible model. Given as such, we propose to model GEI using a shrinkage estimator that combines estimates from Tukey's model and from the saturated interaction model. An adaptive framework is utilized similar to that described by Mukherjee *et al.* [2]. This estimator will shrink the maximum likelihood estimates (MLEs) under a flexible interaction structure toward Tukey's model estimates. The amount of shrinkage is data adaptive, so that in large samples, such estimator is unbiased even if Tukey's assumption is violated. More importantly, when compared to a saturated model, the shrinkage estimator has reduced mean squared error (MSE) for small samples [20]. Although Tukey's model has been used to model GEI or GGI under a generalized linear model setting [15, 18, 19], no prior work has been carried out to data-adaptively combine Tukey's model and saturated interaction model to take advantage of both models for testing GEI. Thus, the shrinkage approach is not only novel for longitudinal data but also a new approach for cross-sectional data.

In Section 2, we introduce notations for GEI models using a mixed-effects model framework. The parameter estimation for Tukey's model with repeated measures data is described in Section 3. In Section 4, we propose a shrinkage estimator and derive its

approximate variance estimate. In Section 5, we summarize the test for interaction corresponding to each method. In Section 6, we evaluate the performance of our proposed methods via simulation studies. In particular, we compare the average performance by generating GEIs with different interaction structures to mimic a hypothetical GEI search study involving multiple genetic markers. In Section 7, we apply the proposed methods to search GEI between 105 single-nucleotide polymorphisms (SNPs) within 22 genes in the iron metabolism pathway and cumulative lead exposure on pulse pressure using the Normative Aging Study (NAS) data. We also test GEI between 27 SNPs and energy intake and intentional exercise on body mass index (BMI) using data from the Multi-Ethnic Study of Atherosclerosis (MESA). These 27 SNPs have been shown to be significantly associated with BMI in previous genome wide association studies (GWAS). In NAS, genes are chosen in relation to the exposure pathway. In MESA, the question is whether the loci identified by GWAS (with marginal effects) modify the effect of certain exposures. Another distinction between the two data examples is that one of the exposure variables considered in MESA, intentional exercise, is a time-varying variable, while the other two, energy intake in MESA and cumulative lead exposure in NAS, are time-invariant (i.e., both are baseline measurements).

2. Model

Let y_{kt} be the value of the t-th repeated measure on a phenotypic response Y corresponding to the k-th individual ($t = 1, \ldots, n_k$, $k = 1, \ldots, N$). Define a mixed-effects model for the $n_k \times 1$ response vector $\boldsymbol{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kn_k})^\top$ such that it is related to an $n_k \times \nu$ matrix of explanatory variables $\boldsymbol{X}_k = (\boldsymbol{x}_{k1}, \boldsymbol{x}_{k2}, \ldots, \boldsymbol{x}_{kn_k})^\top$, with each $\boldsymbol{x}_{kt}$ a $\nu \times 1$ vector associated with y_{kt}, through some nonlinear function f. Namely,

$$\boldsymbol{y}_k = \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k) + \boldsymbol{Z}_k \boldsymbol{b}_k + \boldsymbol{e}_k, \quad (1)$$

where $\boldsymbol{\eta}$ is the p-dimensional vector of fixed effects, $\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)$ is the $n_k \times 1$ mean vector, $\boldsymbol{b}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$ is the q-dimensional vector of random effects, $\boldsymbol{Z}_k$ is the design matrix of size $n_k \times q$ for the random effects satisfying rank($\boldsymbol{Z}_k$) = $q \quad n_k$ for all k, and $\boldsymbol{e}_k = (e_{k1}, \ldots, e_{kn_k})^\top \sim \mathcal{N}(\boldsymbol{0}, \Sigma_k)$ is the n_k-dimensional vector of random errors. The random effects $\boldsymbol{b}_k$ are assumed to be independent of $\boldsymbol{e}_k$. Let $\boldsymbol{V}_k(\boldsymbol{\omega})$ be the variance matrix of $\boldsymbol{y}_k$, $\boldsymbol{V}_k(\omega) = \boldsymbol{Z}_k \boldsymbol{\Psi} \boldsymbol{Z}_k^\top + \sum_k$. Here $\boldsymbol{\omega}$ consists of parameters in $\boldsymbol{\Psi}$ and Σ_k.

We use (1) to model the association between the phenotypic response of interest and genetic and environmental exposure factors. Let G_k be the genotype and E_{kt} be the exposure level for the k-th subject at the t-th measurement, $G_k = i$, $i = 1, 2, \ldots, I$, $E_{kt} = j$, $j = 1, 2, \ldots, J$. Both G_k and E_{kt} are assumed to be categorical variables. Without considering any covariates, the mean structure for y_{kt} under Tukey's model [14] has the following form

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}, \theta, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k = i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \theta \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_i^G \beta_j^E I(G_k = i, E_{kt} = j). \quad (2)$$

Here $\boldsymbol{\eta}$ has two components, $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \theta)^\top$. $\boldsymbol{\beta}$ consists of the intercept β_0, the parameters for genetic main effects, $\boldsymbol{\beta}^G = (\beta_1^G, \ldots, \beta_I^G)^\top$, and exposure main effects, $\boldsymbol{\beta}^E = (\beta_1^E, \ldots, \beta_J^E)^\top$. θ is a scale parameter representing the interaction effect. A saturated interaction model, on the other hand, allows for separate interaction parameters for each GEI configuration:

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k = i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \sum_{i=1}^{I} \sum_{j=1}^{J} \tau_{ij} I(G_k = i, E_{kt} = j), \quad (3)$$

where $\boldsymbol{\tau} = (\tau_{11}, \ldots, \tau_{IJ})^\top$ is the interaction parameter vector with length IJ. Due to the constraints for parameter identifiability, $\sum_i \beta_i^G = \sum_j \beta_j^E = 0$, $\boldsymbol{\beta}^G$ and $\boldsymbol{\beta}^E$ are left with $(I-1)$ and $(J-1)$ independent parameters to be estimated, respectively. Similarly, $\Sigma_i \tau_{ij} = \Sigma_j \tau_{ij} = 0$, so $(I-1)(J-1)$ parameters in $\boldsymbol{\tau}$ are left to be estimated.

3. Parameter Estimation for Tukey's Model with Repeated Measures Data

We describe the estimation strategy for the parameters of Tukey's model. The log-likelihood for the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ is

$$\ell(\boldsymbol{\eta}, \boldsymbol{\omega}, |\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \text{const.} - \frac{1}{2} \sum_{k=1}^{N} \log |V_k(\boldsymbol{\omega})| - \frac{1}{2} \sum_{k=1}^{N} \left\{ [\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)]^\top V_k(\boldsymbol{\omega})^{-1} [\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)] \right\}. \quad (4)$$

Given $V_k(\boldsymbol{\omega})$, maximizing the likelihood is equivalent to minimizing the objective function

$$Q(\boldsymbol{\eta}|\boldsymbol{\omega}) = \sum_{k=1}^{N} [\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)]^\top V_k(\boldsymbol{\omega})^{-1} [\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)] \quad (5)$$

with respect to $\boldsymbol{\eta}$. The solution for $\boldsymbol{\eta}$ is the generalized least squares (GLS) estimator. Since the estimation for fixed effects in Tukey's model does not have a closed-form solution, the iterative linearization method is considered.

The linearization method uses a first-order Taylor series expansion to approximate solutions of a general function by a linear function [21], which has been applied to nonlinear mixed-effects models [22, 23, 24]. Let $\boldsymbol{\eta}^* = \boldsymbol{\eta}^{(\hat{0})} = (\boldsymbol{\beta}^{(\hat{0})\top}, \theta^{(\hat{0})})^\top$ denote the initial estimate of $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \theta)^\top$. The first-order Taylor series expansion of $\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)$ about $\boldsymbol{\eta} = \boldsymbol{\eta}^*$ is

$$\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k) \approx \boldsymbol{f}(\boldsymbol{\eta}^*, \boldsymbol{X}_k) + \boldsymbol{D}_k^* (\boldsymbol{\eta} - \boldsymbol{\eta}^*), \quad (6)$$

where $\boldsymbol{D}_k^*$ is an $n_k \times p$ matrix $\boldsymbol{D}_k^{*\top} = \boldsymbol{D}_k^\top(\boldsymbol{\eta}^*) = \{\partial \boldsymbol{f}(\boldsymbol{\eta})/\partial \eta_1, \ldots, \partial \boldsymbol{f}(\boldsymbol{\eta})/\partial \eta_p\}|_{\boldsymbol{\eta}^*}$. Initial values of $\boldsymbol{\eta}^*$ can be obtained by fitting a saturated interaction model (via standard linear mixed effects model) and using the main effect estimates as $\boldsymbol{\beta}^*$. After removing main effects, the residuals can then be regressed on the product term $\beta_i^{G*} \beta_j^{E*}$ (without intercept) to obtain θ^*. The mean function of Tukey's model for the k-th subject at the t-th measurement is

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) \approx f(\boldsymbol{\eta}^*, \boldsymbol{x}_{kt})$$
$$+ (\beta_0 - \beta_0^*) + \sum_i \sum_j [(1$$
$$+ \theta^* \beta_j^{E*})(\beta_i^G$$
$$- \beta_i^{G*}) + (1 + \theta^* \beta_i^{G*})(\beta_j^E$$
$$- \beta_j^{E*}) + \beta_i^{G*}\beta_j^{E*}(\theta$$
$$- \theta^*)] I(G_k$$
$$= i, E_{kt} = j),$$

where

$$f(\boldsymbol{\eta}^*, \boldsymbol{x}_{kt}) = \beta_0^* + \sum_i \beta_i^{G*} I(G_k = i) + \sum_j \beta_j^{E*} I(E_{kt} = j) + \theta^* \sum_i \sum_j \beta_i^{G*} \beta_j^{E*} I(G_k = i, E_{kt} = j).$$
Following (1), the expansion in (6) yields the approximation

$$\boldsymbol{y}_k = \boldsymbol{f}(\boldsymbol{\eta}^*, \boldsymbol{X}_k) + \boldsymbol{D}_k^*(\boldsymbol{\eta} - \boldsymbol{\eta}^*) + \boldsymbol{Z}_k \boldsymbol{b}_k + \boldsymbol{e}_k,$$

which can be expressed as a linear model

$$\boldsymbol{y}_k^* = \boldsymbol{D}_k^* \boldsymbol{\eta} + \boldsymbol{Z}_k \boldsymbol{b}_k + \boldsymbol{e}_k, \quad (7)$$

where $\boldsymbol{y}_k^* = \boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}^*, \boldsymbol{X}_k) + \boldsymbol{D}_k^* \boldsymbol{\eta}^*$. Then the GLS estimator for $\boldsymbol{\eta}$ is given by

$$\hat{\boldsymbol{\eta}}_{GLS} = \left(\sum_{k=1}^N \boldsymbol{D}_k^{*\top} \hat{\boldsymbol{V}}_k^{*-1} \boldsymbol{D}_k^* \right)^{-1} \sum_{k=1}^N \boldsymbol{D}_k^{*\top} \hat{\boldsymbol{V}}_k^{*-1} \boldsymbol{y}_k^*, \quad (8)$$

where $\hat{\boldsymbol{V}}_k^*$ is the assumed covariance matrix of $\boldsymbol{y}_k^*$ evaluated at $\boldsymbol{\omega} = \boldsymbol{\omega}^*$. When $\boldsymbol{\eta}$ and $\boldsymbol{\omega}$ are unknown, a common strategy is to replace $\boldsymbol{V}(\boldsymbol{\omega})$ with a consistent estimate and minimize the corresponding weighted sum of squares to yield an initial estimate of $\boldsymbol{\eta}$. The MLE of $\boldsymbol{\omega}$ is obtained by maximizing (4) with respect to $\boldsymbol{\omega}$, after $\boldsymbol{\eta}$ is replaced by the estimate in (8).

This iteratively reweighted generalized least-squares (IRGLS) algorithm involves iterations between [a] Taylor series linearization – given the w-th iterates $\hat{\boldsymbol{\eta}}^{(w)}$ and $\hat{\boldsymbol{\omega}}^{(w)}$, construct $\boldsymbol{D}_k^{(w)} = \boldsymbol{D}(\hat{\boldsymbol{\eta}}^{(w)})$ and $\hat{\boldsymbol{r}}_k^{(\omega)} = \boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}^{(w)}, \boldsymbol{X}_k) + \boldsymbol{D}_k^{(w)} \hat{\boldsymbol{\eta}}^{(w)}$ to yield a pseudo model that is of the form of (7) – and [b] updating estimates $\hat{\boldsymbol{\eta}}^{(w+1)}$ in (8) and $\hat{\boldsymbol{\omega}}^{(w+1)}$. Steps [a] and [b] are repeated until a convergence criterion is achieved.

The linearization method provides an easy calculation for nonlinear models by translating the nonlinear estimation problem into a linear model. Only the first-order derivatives are required. Though the assumption of normality is not required for estimates from this IRGLS procedure, minimizing the objective function (5) is equivalent to maximizing the joint log-likelihood function of $\boldsymbol{y}_k$ in (4). Hence, this procedure yields MLEs [25]. Vonesh *et al.* [26]

argued that the IRGLS estimator is consistent and asymptotically normal even when the variance-covariance structure is misspecified if the mean function $f(\eta, X_k)$ is correctly specified. Our experience is that the proposed estimation algorithm for Tukey's model converges relatively fast and the final estimates are insensitive to initial values. Nevertheless, seriously slow convergence or possibly non-convergence could occur when one or both of the main effects are truly absent, a situation where θ is not identifiable.

4. Shrinkage Estimator

We now construct a shrinkage estimator for interaction that is a weighted average of the estimators from Tukey's model and a saturated interaction model. Denote the interaction parameters to be estimated for an $I \times J$ GEI table by $\boldsymbol{\tau} = (\tau_{11}, \tau_{21}, \ldots, \tau_{(I-1)1}, \tau_{12}, \ldots, \tau_{(I-1)(J-1)})^{\top}$. Let $\boldsymbol{\tau}_{tuk}$ and $\boldsymbol{\tau}_{sat}$ be the asymptotic limits of the estimator of $\boldsymbol{\tau}$ from Tukey's model and saturated interaction model, respectively, each being a length-$(I-1)(J-1)$ vector. When the true model is a Tukey's one-df model, we have $\boldsymbol{\tau}_{tuk} - \boldsymbol{\tau}_{sat} = \boldsymbol{\delta}(\text{say}) = \mathbf{0}$. To relax the model assumption, let $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta})$. A conservative estimate of $\boldsymbol{\Theta}$ is given by $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top}$, where $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\tau}}_{tuk} - \hat{\boldsymbol{\tau}}_{sat}$ and $\hat{\boldsymbol{\tau}}_{tuk} = \hat{\theta}(\hat{\beta}_1^G \hat{\beta}_1^E, \hat{\beta}_2^G \hat{\beta}_1^E, \ldots, \hat{\beta}_{I-1}^G \hat{\beta}_{J-1}^E)^{\top}$. We define $\boldsymbol{B} = \hat{\boldsymbol{V}}_{\tau}(\hat{\boldsymbol{V}}_{\tau} + \hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top})^{-1}$, where $\hat{\boldsymbol{V}}_{\tau}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\tau}}_{sat}$. Then the proposed shrinkage estimator for $\boldsymbol{\tau}$ is given by

$$\hat{\boldsymbol{\tau}}_{shk} = \hat{\boldsymbol{\tau}}_{sat} + \boldsymbol{B}(\hat{\boldsymbol{\tau}}_{tuk} - \hat{\boldsymbol{\tau}}_{sat}), \quad (9)$$

where $\hat{\boldsymbol{\tau}}_{tuk}$ and $\hat{\boldsymbol{\tau}}_{sat}$ are MLEs from (2) and (3), respectively.

The shrinkage factor $\boldsymbol{B}$ in (9) determines the amount of shrinkage of $\hat{\boldsymbol{\tau}}_{sat}$ toward $\hat{\boldsymbol{\tau}}_{tuk}$. As $\hat{\boldsymbol{\delta}} \to \mathbf{0}$ and $\boldsymbol{B} \to \boldsymbol{I}$, $\hat{\boldsymbol{\tau}}_{shk} \to \hat{\boldsymbol{\tau}}_{tuk}$ (data are indicative of a Tukey's interaction structure). On the other hand, as the bias of Tukey's model estimator $\hat{\boldsymbol{\delta}}$ increases, the largest eigenvalue of $\boldsymbol{B}$ goes to 0 and $\hat{\boldsymbol{\tau}}_{shk} \to \hat{\boldsymbol{\tau}}_{sat}$ (data are not in favor of Tukey's form of interaction). Now express the shrinkage estimator in (9) as

$$\hat{\boldsymbol{\tau}}_{shk} = \hat{\boldsymbol{\tau}}_{sat} + \hat{\boldsymbol{V}}_{\tau} \left(\hat{\boldsymbol{V}}_{\tau}^{-1} - \frac{\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}} \right) \hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\tau}}_{sat} + \hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}} \left(\frac{\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}} \right).$$

When data are under Tukey's model, $\hat{\boldsymbol{\delta}} \to \mathbf{0}$ as $N \to \infty$. When data are not under Tukey's model, the largest eigenvalue of $\hat{\boldsymbol{V}}_{\tau}$ goes to 0 and $\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}} \to \infty$ as $N \to \infty$. So, the term $(\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}})/(1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}})$ converges to 1. This indicates that $\hat{\boldsymbol{\tau}}_{shk}$ is asymptotically equivalent to $\hat{\boldsymbol{\tau}}_{sat}$, which is an unbiased estimator of $\boldsymbol{\tau}$. But with moderate sample size, $\hat{\boldsymbol{\delta}}$ creates a small bias in $\hat{\boldsymbol{\tau}}_{shk}$ that can be traded for a larger decrease in variance, leading to an improvement in finite sample MSE [2]. In addition, when main effects are not present, the shrinkage estimator will guard against the instability of parameter estimates under Tukey's model by shrinking $\hat{\boldsymbol{\tau}}_{shk}$ toward $\hat{\boldsymbol{\tau}}_{sat}$.

4.1. Variance Estimation for the Shrinkage Estimator

We proceed to estimate the covariance matrix for $\hat{\boldsymbol{\tau}}_{shk}$. As a result of asymptotic equivalence of $\hat{\boldsymbol{\tau}}_{shk}$ and $\hat{\boldsymbol{\tau}}_{sat}$, the covariance matrix for $\hat{\boldsymbol{\tau}}_{sat}$ can be used as an estimator for the covariance matrix of $\hat{\boldsymbol{\tau}}_{shk}$ in large samples. Since this estimator is often too conservative in finite samples, we develop an approximate covariance matrix estimator for $\hat{\boldsymbol{\tau}}_{shk}$ using the delta method.

Define $\hat{\boldsymbol{\phi}} = (\boldsymbol{\tau}_{sat}^{\top}, \hat{\boldsymbol{\eta}}_{tuk}^{\top})^{\top}$ as the MLEs under a saturated form of interaction and Tukey's model with $\hat{\boldsymbol{\eta}}_{tuk} = (\hat{\beta}_1^G, \ldots, \hat{\beta}_{I-1}^G, \hat{\beta}_1^E, \ldots, \hat{\beta}_{J-1}^E, \hat{\theta})^{\top}$. Further define $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\tau}}_{sat}, \hat{\boldsymbol{\tau}}_{tuk})^{\top} = \boldsymbol{h}(\hat{\boldsymbol{\varphi}})$ such that $\hat{\boldsymbol{\tau}}_{shk} = \boldsymbol{g}(\hat{\boldsymbol{\xi}}) = \boldsymbol{g}(\boldsymbol{h}(\hat{\boldsymbol{\varphi}}))$, where $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{g}(\hat{\boldsymbol{\xi}})$ have $2(I-1)(J-1)$ and $(I-1)(J-1)$ elements, respectively. We first derive the joint distribution of the components in $\hat{\boldsymbol{\varphi}}$. Let $\mathcal{I}$ be the information matrix with dimension $(I-1)(J-1) \times (I-1)(J-1)$ and ℓ be the log-likelihood corresponding to a saturated interaction model (3). Let $\mathcal{I}_0$ be the information matrix with dimension $(I+J-1) \times (I+J-1)$ and ℓ_0 be the log-likelihood for Tukey's model (2). By the consistency of $\hat{\boldsymbol{\varphi}}$, the MLE $\hat{\boldsymbol{\tau}}_{sat}$ has an asymptotic linear representation

$$\sqrt{N}(\hat{\boldsymbol{\tau}}_{sat} - \boldsymbol{\tau}) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathcal{I}^{-1} \dot{\ell}_k + o_p(1) \text{ as } N \to \infty, \text{ where } \dot{\ell}_k = \partial \ell(\boldsymbol{X}_k)/\partial \boldsymbol{\tau}.$$

Similarly,

$$\sqrt{N}(\hat{\boldsymbol{\eta}}_{tuk} - \boldsymbol{\eta}) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathcal{I}_0^{-1} \dot{\ell}_{0k} + o_p(1) \text{ as } N \to \infty, \text{ where } \dot{\ell}_{0k} = \partial \ell_0(\boldsymbol{X}_k)/\partial \boldsymbol{\eta}_{tuk}.$$

Denote the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\varphi}}$ by $\Sigma_{\hat{\boldsymbol{\varphi}}}$. Then by multivariate Taylor series expansion, the variance-covariance matrix of $\hat{\boldsymbol{\xi}} = \boldsymbol{h}(\hat{\boldsymbol{\varphi}})$ is approximated by

$$\hat{\sum}_{\hat{\boldsymbol{\xi}}} \approx \{\nabla \boldsymbol{h}(\hat{\boldsymbol{\phi}})\}^{\top} \hat{\sum}_{\hat{\boldsymbol{\phi}}} \nabla \boldsymbol{h}(\hat{\boldsymbol{\phi}}),$$

where $\nabla \boldsymbol{h} = \partial \boldsymbol{h}/\partial \boldsymbol{\varphi}$ is the gradient matrix of $\boldsymbol{h}$ evaluated at $\hat{\boldsymbol{\varphi}}$. Finally, the variance-covariance matrix of $\hat{\boldsymbol{\tau}}_{shk}$ is approximated by applying the delta method:

$$\hat{\sum}_{\hat{\boldsymbol{\tau}}_{shk}} = \hat{\text{cov}}(\hat{\boldsymbol{\tau}}_{shk}) = \hat{\text{cov}}(\boldsymbol{g}(\hat{\boldsymbol{\xi}})) \approx \{\nabla \boldsymbol{g}(\hat{\boldsymbol{\xi}})\}^{\top} \hat{\sum}_{\hat{\boldsymbol{\xi}}} \nabla \boldsymbol{g}(\hat{\boldsymbol{\xi}}), \quad (10)$$

where $\nabla \boldsymbol{g} = \partial \boldsymbol{g}/\partial \boldsymbol{\xi}$ evaluated at $\hat{\boldsymbol{\xi}}$ (refer to the supporting information for $\nabla \boldsymbol{h}(\hat{\boldsymbol{\varphi}})$ and $\nabla \boldsymbol{g}(\hat{\boldsymbol{\xi}})$.) Comparing $\Sigma_{\hat{\boldsymbol{\tau}}_{shk}}$ to the empirical estimate of variance-covariance matrix through simulations, we found that variance components can be estimated very well by $\Sigma_{\hat{\boldsymbol{\tau}}_{shk}}$ but not necessarily the covariance. Either a small variance for the random measurement errors or a large sample size is needed to obtain accurate estimates of covariance terms (see Table 1 in the supporting information). Since the magnitudes of covariance estimates are smaller

compared to the variance estimates, the influence of covariance estimates on the Wald test statistic is expected to be small. Thus, the proposed shrinkage test (see below) is still an approximately valid test with conservative Type 1 error rates.

5. Tests for Interaction Effects

We are interested in testing the null hypothesis of no interaction effects $H_0 : \boldsymbol{\tau} = \mathbf{0}$ versus $H_1 : \boldsymbol{\tau} \neq \mathbf{0}$. For Tukey's model, it is equivalent to $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. A likelihood ratio test (LRT) statistic is given by $T_L = -2(l_0 - l_1)$, where l_0 and l_1 are the maximized log-likelihoods obtained under H_0 and H_1, respectively. Under regularity conditions, $T_L \sim \chi_1^2$ for Tukey's model and $T_L \sim \chi_{(I-1)(J-1)}^2$ for saturated model under H_0 for large samples. Based on (9) and [20], the limiting distribution of the shrinkage estimator is technically not normal. The simulation results, however, reveal that this estimator is well approximated by a normal density and the amount of departure from normality is small (see Figure 1 in the supporting information). Hence, the Wald test is used as an approximate test for interaction. The test statistic for $H_0 : \boldsymbol{\tau} = \mathbf{0}$ is given by $\tilde{T}_W = \hat{\boldsymbol{\tau}}_{shk}^\top \hat{\sum}_{\hat{\boldsymbol{\tau}}_{shk}}^{-1} \hat{\boldsymbol{\tau}}_{shk}$, where $\hat{\sum}_{\hat{\boldsymbol{\tau}}_{shk}}$ can be found in (10). $\tilde{T}_W$ approximately follows a χ^2 with $df = (I - 1)(J - 1)$ under H_0 (see Figure 2 in the supporting information).

6. Simulation Study

6.1. Settings for Evaluation of Test Properties for a Single GEI Test

We investigated the Type I error and power properties of the following three test procedures for interaction: the LRT under Tukey's model of interaction, the Wald test using the proposed adaptive shrinkage estimator, and the LRT using a saturated interaction model. Two null hypotheses of no interactions were considered: (i) the genetic main effects were present (additive) and (ii) the genetic main effect were absent (null). The main effects of the exposure were always present in our simulations to represent a study looking for genetic modification effects on an established phenotype-exposure association. For these comparisons, we used 3×3 table settings for GEI with N=1200. The number of repeated measurements per subject was generated from a multinomial distribution similar to the example data: $n_{ijk} \in \{2, 3, 4, 5, 6\}$, $\boldsymbol{n} = \{n_{ijk} : 1 \leq k \leq N_{ij}, 1 \leq i \leq I, 1 \leq j \leq J\} \sim \text{Mult}(N, \boldsymbol{p})$, $\boldsymbol{p} = (0.15, 0.2, 0.3, 0.2, 0.15)$, which implies that dropouts are missing completely at random. Data were simulated under a first-order autoregressive (AR-1) correlation structure: $\sigma^2 \rho^{|t-t'|}$ for the (t, t')-th element in Σ_k ($\sigma^2 = 4, 8$ and $\rho = 0.7$). Additionally, the test properties were evaluated under misspecification of correlation structure. Again, data were still generated under the AR-1 correlation structure but were analyzed using a compound symmetric covariance structure. A total of 1000 datasets were generated for each setting. Type I error and power were estimated by the sample proportions of null hypothesis being rejected under various simulation settings.

In the 3×3 GEI table settings, three genotype categories were considered for G with minor allele frequency 0.4 and following the Hardy-Weinberg equilibrium. An environmental exposure with three categories (with probabilities 0.25, 0.25, and 0.50) was considered. Cell

means for all GEI configurations were first generated under a pre-specified interaction model. Given a mean and covariance structure, the vector of observations per individual were generated from a multivariate normal distribution. In addition to Tukey's and saturated models, we considered simulations under additive main effects and multiplicative interaction (AMMI) models [27, 28]. AMMI models are a class of interaction models that have a flexible structure, which essentially entails a singular value decomposition (SVD) of the cell residual matrix after removing the additive main effects. Following the notations in (2), the mean structure for y_{kt} under an AMMI model is given by

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\beta, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k=i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt}=j) + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{M} d_m \alpha_{im} \gamma_{jm} I(G_k=i, E_{kt}=j).$$

The m-th interaction factor is subject to the constraints $\sum_{i=1}^{I} \alpha_{im}^2 = \sum_{j=1}^{J} \gamma_{jm}^2 = 1$ and

$\sum_{i=1}^{I} \alpha_{im} = \sum_{j=1}^{J} \gamma_{jm} = 0$, as well as the $2(M-1)$ orthogonality restrictions $\Sigma_i \, a_{im} a_{im'} = \Sigma_i \, \gamma_{jm} \gamma_{jm'} = 0$ for $m \neq m'$. Specifically, AMMI models with $M = 1$ (AMMI1) were considered in the simulation as an intermediate model between Tukey and saturated model. AMMI2 would be equivalent to a saturated interaction model in the 3×3 table settings. We compared test performance under AMMI1 models because Tukey's test may not be capable of capturing interaction of AMMI1 form. Though AMMI1 is nested within the saturated interaction model, the test based on a saturated interaction model may not have as much power to detect the interaction.

6.2. Settings for Assessment of Average Performance for Multiple GEI Tests

When GEI tests are conducted across a moderately large number of SNPs within several gene regions, the average performance of each method over many GEI tests is of particular interest rather than a single specific GEI test. As such, we assessed the Type I error and power of the tests for interaction using Tukey's model, saturated interaction model, and the proposed shrinkage estimator, averaged over a set of genetic markers. We based our simulation studies on the setting of the NAS data example where the candidate genes were chosen based on some pathway analysis. For each dataset, one exposure factor and 100 independent SNPs (without LD) were generated, with the minor allele frequencies ranging from 0.3 to 0.5. (Simulations with a wider range of minor allele frequencies can be found in the supporting information.) The exposure had five categories, each with probability 0.2. Thus, a 3×5 table was constructed for each GEI test.

We considered two simulation schemes for multiple GEI tests: (i) 100 marginal models, $Y_i|G_i, E$, $i = 1, ..., 100$, were generated with a common E for each subject, and (ii) a joint multivariate model, $Y|G_1, G_2, ..., G_{100}, E$, was generated. In both (i) and (ii), 15 out of 100 SNPs were assigned to have GEI effects on Y. Another five SNPs were generated to have only additive main effects on Y. The rest 80 SNPs were not associated with Y. The simulation design represents a study where GEI over multiple SNPs are being tested, the

majority of SNPs do not have GEI effects and only a relatively small number of SNPs exhibit GEI effects.

To assess the sensitivity of tests in response to the underlying composition of different interaction models, we created three scenarios by assigning each of the 15 GEI to have either a Tukey's or a saturated form of interaction: Scenario (A): all 15 were of Tukey's form of interaction; Scenario (B): 10 were of Tukey's form, and 5 had saturated interaction structures; Scenario (C): 10 had saturated interaction structures, and 5 were of Tukey's form. For example, the mean function of the simulation model for subject k under scenario (B) in simulation scheme (ii), following the notations in (3), is given by

$$
\begin{aligned}
f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = & \beta_0 + \sum_{j=1}^{J} \beta_j^E I(E_{kt}=j) + \sum_{i=1}^{I} \sum_{j=1}^{J} \left\{ \sum_{s=1}^{10} [\beta_i^{Gs} I(G_{sk}=i) + \theta^s \beta_i^{Gs} \beta_j^E I(G_{sk}=i, E_{kt}=j)] \right. \\
& \left. + \sum_{s=11}^{15} [\beta_i^{Gs} I(G_{sk}=i) + \tau_{ij}^s (G_{sk}=i, E_{kt}=j)] \right\} + \sum_{i=1}^{I} \sum_{s=15}^{20} \beta_i^{Gs} I(G_{sk}=i),
\end{aligned}
$$

where β_i^{Gs} represents the genetic main effect of the i-th genotype from the s-th SNP, G_{sk} is the genotype of the s-th SNP for the k-th subject, and θ^s and τ_{ij}^s are the interaction parameter corresponding to the s-th SNP. An individual-level outcome Y with repeated measures were generated for 1000 subjects in each simulation using (1) with $\boldsymbol{e}_k \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_{n_k})$, $\boldsymbol{b}_k = b_k \mathbf{1}_{n_k}$, $b_k \sim \mathcal{N}(0, \sigma_b^2)$. We set $\sigma_b^2 = 2.8$, and $\sigma_e^2 = 1.2$. The number of repeated measurements per subject was generated using the same multinomial distribution described previously.

The average performance for each test procedure was quantified by true positive rate (TPR) and false positive rate (FPR). The TPR is defined as the proportion of interactions detected in the 15 simulated SNPs with GEI associations. The FPR is the proportion of interactions detected among the 85 simulated SNPs without GEI effects. The TPR and FPR were then averaged over 10,000 simulation datasets. To control the family-wise error rate (FWER), the significance level was adjusted according to the total number of SNPs (i.e., number of GEI tests) using Bonferroni correction, $\alpha^* = 0.05/100 = 5 \times 10^{-4}$.

6.3. Power and Type I Error

The upper panel of Table 1 shows the power and Type I error of tests using Tukey's, the saturated model, and the shrinkage estimator for GEI. In general, the saturated interaction model has less power to detect interactions when the true interaction has a Tukey's form. For example, the LRT for Tukey's form of interaction has power 0.76 for $\sigma^2 = 4$, while the saturated model has a power of 0.54. On the other hand, when the true interaction has a saturated form, Tukey's model can hardly detect the interaction effects. The saturated model has a power of 0.81 for $\sigma^2 = 4$, but Tukey's model using the LRT only has power 0.09. Under both situations, the interaction test using the shrinkage estimator has power 0.69. When the true interaction has an AMMI1 form, the saturated interaction and the shrinkage estimator can detect 82% and 72% of interactions, respectively, but Tukey's model can only detect 30% of interactions. The Type I error rates are maintained at the nominal level for all testing procedures under additive models except the Wald test using the shrinkage estimator

being a slightly conservative test. However, both Tukey's test and the shrinkage estimator have inflated Type I error under the completely null model when one of the main effects is not present.

When the within-subject correlation structure is misspecified (lower panel of Table 1), the patterns of power comparison are similar to the upper panel. Under the null hypothesis of an additive model where both main effects are present, the Type I error rates for the two LRTs are still maintained at the 0.05 level when $\sigma^2 = 4$ but are inflated when $\sigma^2 = 8$. Only the proposed Wald test using the shrinkage estimator maintains the nominal level of Type I erorr. Under the null that genetic main effects are absent, the Type I error is no longer maintained at 0.05 for all of the tests.

6.4. Average Performance for Multiple GEI Tests

The upper panel of Table 2 shows the average performance of the three GEI tests for marginal models under three scenarios. Under scenario (A) where all 15 simulated GEI are of Tukey's form, the LRT using Tukey's model has a TPR of 0.72, whereas the saturated model has a TPR of 0.43. Under scenario (B) where 2/3 of the simulated GEI are of Tukey's form, the LRT using Tukey's model and the saturated interaction test have comparable performance. Under scenario (C) where 2/3 of the interactions are of saturated forms, the Wald test using the shrinkage estimator and the saturated interaction tests have comparable performance, but the TPR for the LRT using Tukey's model is substantially lower. The FPRs are maintained at the nominal level for the tests using a saturated model and slightly inflated for the shrinkage estimator. However, the LRT for Tukey's model has the highest FPR.

The lower panel of Table 2 shows the results of a multivariate model (single outcome) from 100 simulated GEI. The LRT using a saturated interaction form yields relatively low TPRs. The test based on the shrinkage estimator still maintains at the same level of TPR across scenarios. In summary, the GEI test using the shrinkage estimator has the most robust average performance with respect to various GEI structures compared to the tests using Tukey's and saturated interaction models.

7. Application

7.1. Normative Aging Study (NAS)

The Normative Aging Study (NAS) is a multidisciplinary longitudinal study initiated by the U.S. Veterans Administration in 1963 to investigate the effects of aging on various health outcomes [29]. We focus on pulse pressure (PP), which is an important risk factor for heart disease [30]. Several studies have indicated a relationship between iron deficiency and increased lead absorption [31, 32], and increased cumulative lead exposure has been shown to be associated with elevated PP [33]. Thus, it may be reasonable to hypothesize that genes responsible for iron metabolism could potentially alter lead absorption and modify the effect of lead exposure on PP. The objective of this pathway-driven GEI study was to test the GEI between cumulative lead exposure and the iron metabolic genes on PP.

Zhang *et al.* [34] observed a significant interaction between polymorphisms in the hemochromatosis (*HFE*) gene (rs1799945) and cumulative lead exposure on PP. We revisited the study to include 105 SNPs in 22 genes with minor allele frequency>0.1 in the iron metabolic pathway to test for GEI using the proposed shrinkage estimation framework. Candidate genes were chosen based on *a priori* knowledge of iron metabolism and previous studies on iron-related genes [35, 36]. We analyzed 729 participants from a subset of the NAS data who were successfully genotyped for the iron metabolism genes and had baseline measurements of cumulative lead concentrations (measured at the tibia bone and patella bone). The majority (97%) of the participants were Caucasian. The average age was 66.37±7.12 (range 48–93) at the time of bone lead measurement. Since 1991, blood pressure had been measured every 3–5 years until 2011 with a median follow-up time of 12 years. More than 94% of subjects had repeated measurements of blood pressure, and over 48% of them had at least four measurements during the study period contributing to a total of 3013 observations (see Table 7 in the supporting information).

Each of the 105 SNPs had three possible genotypes (homozygous wild-type, heterozygous, and homozygous mutant). For illustration purposes, we categorized bone lead concentrations into three groups – Low: 15, Medium: (15, 25], and High: >25 μg/g for the tibia bone lead and Low: 20, Medium: (20, 32], and High: >32 μg/g for the patella bone lead. We used Tukey's model, saturated interaction model, and the shrinkage approach to model the GEI structures for each *SNP×Lead* interaction. Covariates in the model included baseline age, time since baseline, and squared time. According to the Akaike information criterion (AIC) for model fit, we chose a random-intercept mixed-effects model for analysis given by $y_k = f(\pmb{\eta}, \pmb{X}_k) + b_k\mathbf{1}_{n_k} + \pmb{e}_k$, where $b_k \sim \mathcal{N}(0, \sigma_b^2)$, $\pmb{e}_k \sim \mathcal{N}(0, \sigma_e^2\pmb{I}_{nk})$.

Given that these SNPs are located in a small number of genomic regions, they are in close proximity to each other and thus may exhibit LD. To control for the FWER while accounting for the potentially correlated SNPs in the multiple testing procedure, we adjusted the significance level according to the effective number of independent tests (denoted by M_{eff}) using the simple *M* method [37]. This method involves first estimating the correlation matrix among the 105 SNPs by the composite LD, calculating the corresponding eigenvalues, $\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_{105}$, and then finding M_{eff} through principal component analysis: $\sum_{s=1}^{M_{\mathrm{eff}}} \lambda_s / \sum_{s=1}^{105} \lambda_s > C$. We chose $M_{\mathrm{eff}} = 89$ so that the corresponding eigenvalues explained at least $C = 99.5\%$ of the variation for the SNP data. Thus, the adjusted significance level was $a^* = 0.05/M_{\mathrm{eff}} = 0.05/89 = 5.6 \times 10^{-4}$.

Table 3 lists the smallest *p*-values of GEI tests for the three top-ranked SNPs by using Tukey's model, the proposed shrinkage estimator, and saturated interaction model within iron gene regions in the NAS data. The Wald test via the shrinkage estimator yielded the smallest *p*-values across all top ranked SNPs listed in the table (and three of which reached statistical significance), compared to Tukey's and saturated interaction models. For tibia bone lead, we found a significant modifying effect of SNPrs1799945 in the *HFE* gene using the shrinkage estimator ($p = 1 \times 10^{-4}$). For the wild-type participants, mean PP remained nearly unchanged between the High and the Low tibia lead groups. In contrast, mean PP was estimated to be 20.35 mmHg (95% CI = [14.53, 26.17]) higher for the High tibia lead group

than the Low tibia lead group among the homozygous mutant carriers. The results replicate the findings in Zhang *et al.* [34] that the positive association between PP and lead exposure was strongest among *HFE* homozygous mutant carriers. For patella bone lead, significant modifying effects of SNP rs17484524 in the IREB2 (iron-responsive element binding protein 2) gene ($p = 3 \times 10^{-4}$) and SNP rs7165535 in the B2M (beta-2-microglobulin) gene ($p = 4 \times 10^{-4}$) were detected using the Wald test based on the shrinkage estimator (but were not captured by the LRTs using Tukey's or saturated interaction model). For the wild-type and the heterozygous mutant participants, higher lead levels corresponded to higher mean PP (the estimated difference in mean PP between High and Low patella lead groups ranged from 3.12 to 4.32 mmHg at both SNPs). However, mean PP was estimated to be 3.90 (95% CI = [1.45, 6.35]) and 7.73 (95% CI = [1.88, 13.58]) mmHg *lower* for the High lead group than the Low lead group among the homozygous mutant carriers at SNP rs17484524 in the IREB2 gene and SNP rs7165535 in the B2M gene, respectively. As such, the two homozygous mutant genotypes may indicate protective effects (i.e., preventing PP from elevating with increased lead exposure).

7.2. Multi-Ethnic Study of Atherosclerosis (MESA)

The Multi-Ethnic Study of Atherosclerosis (MESA) is a longitudinal study to investigate characteristics related to progression of subclinical to clinical cardiovascular disease [38]. More than 6,800 men and women aged 45–84 years were recruited from six U.S. communities. Participants had a baseline examination (exam 1) in 2000–2002 and three additional follow-up examinations 18–24 months apart (exams 2–4). We aimed to explore GEI effects on BMI in the four race groups: Caucasians (*N*=2526), Chinese (*N*=775), African Americans (*N*=1611), and Hispanics (*N*=1449). Most (84%) of the participants had four BMI measurements, and over 92% had at least two measurements during the study period from 2000 to 2007 (see Table 3 in the supporting information). A total of 27 SNPs that have demonstrated significant and replicated evidence of marginal association with BMI were selected as the candidate SNPs [39]. The environmental exposures of interest were energy intake, measured at exam 1, and total intentional exercise, measured at exams 1–3. Both exposure variables were categorized into five groups: 0, (0, 7], (7, 14], (14, 28], >28 (MET-hr/week) for total intentional exercise and <1000, (1000, 1300], (1300, 1600], (1600, 2000], >2000 (kcal/day) for energy intake.

We applied Tukey's model, saturated interaction model, and the shrinkage test to examine the GEI structure for each SNP×Energy Intake and SNP×Exercise interaction. Covariates in the model included age at the time of data collection (centered), squared age, gender, having a college degree, household income, and the exposure variable (either intentional exercise or energy intake). We also accounted for population stratification by including the first two principal components. Except age, BMI, and intentional exercise that changed with time, all other variables were time-invariant. We chose an unstructured covariance matrix for this analysis based on AIC. A random gender effect was added to allow men and women to have different variances in BMI. Let $\mathbf{F} = \mathbf{1}_{n_k}$ for women and $\mathbf{F} = \mathbf{0}_{n_k}$ for men. The analysis model is given by $\boldsymbol{y}_k = f(\boldsymbol{\eta}, \boldsymbol{X}_k) + \mathbf{F}_k b_k + \boldsymbol{e}_k$, where $b_k \sim (0, \sigma_b^2)$, $\boldsymbol{e}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$. We first analyzed data by race group (see Table 8 in the supporting information) and then applied Fisher's method [40] to combine four race groups into a single meta-analysis *p*-value for each SNP. Not

every race group allowed for GEI tests across all 27 SNPs because of small sample size in certain GEI configurations. The df for deriving the combined *p*-values was based on the number of available race groups. The adjusted *p*-value to control for the FWER was set at $0.05/27 = 0.0019$.

Table 4 lists the combined *p*-values for significant SNPs using the three interaction tests. For the association of energy intake with BMI, significant modifying effect of SNP rs543874 on the *SEC16B* gene was observed using all three tests. SNP rs1558902 within the *FTO* gene was detected by Tukey's model ($p = 4.8 \times 10^{-5}$) and the shrinkage test ($p = 7.4 \times 10^{-4}$). SNP rs10767664 (on the *BDNF* gene) was also detected by Tukey's model ($p = 1.2 \times 10^{-3}$). For the association between intentional exercise and BMI, we found significant modifying effect of SNP rs206936 within the *NUDT3* and *HMGA1* genes using Tukey's model ($p = 1.4 \times 10^{-4}$). Overall, only one interaction was detected by a standard saturated interaction model used in the current practice. Both the examples illustrate the utility of enhancing power of a test for interaction by leveraging Tukey's model. The shrinkage estimator also offers protection against false positive. The findings require further replication studies.

8. Discussion

We proposed a novel adaptive shrinkage estimator that combines estimates from Tukey's one-df model and a saturated interaction model for GEI effects. The shrinkage estimator shrinks the MLEs under a general, saturated interaction structure toward Tukey's one-df model estimator that allows for data-adaptive relaxation of the structural assumption in Tukey's product form.

The unique simulation setting of multiple GEI tests represents the search for GEI over many candidate SNPs with different interaction patterns. The results indicate that the test based on the shrinkage estimator can be considered as a robust and unified approach for interaction detection. More importantly, the shrinkage method not only can be applied to the context of GEI or GGI detection but also can be extended to any two-way table.

We evaluated MSE and bias of these estimators of interaction effects through simulations (Table 2 in the supporting information). The performance of the shrinkage estimator was compared with the MLE under a general saturated interaction model using the ratio of MSE,

$$\hat{E}\left\{\sum_i\sum_j(\hat{\tau}_{shk_{ij}} - \tau_{ij})^2\right\} / \hat{E}\left\{\sum_i\sum_j(\hat{\tau}_{sat_{ij}} - \tau_{ij})^2\right\}.$$

Based on simulation results, the ratio is uniformly less than 1, suggesting an efficiency advantage for the shrinkage estimator via bias-variance trade-off. In our simulation studies, we noted that the Wald test using the shrinkage estimator is slightly conservative, so the small bias of the shrinkage estimator in finite samples does not lead to inflated Type I error. In addition, we compared the shrinkage estimates of interaction parameters using only the diagonal elements of *B* (i.e., scalar shrinkage) versus using the whole *B* matrix (i.e., multivariate shrinkage). We found that multivariate shrinkage is required under certain

situations (see Table 3 in the supporting information). Chen *et al.* [20] proposed both multivariate and scalar shrinkage estimators in case-control studies, and they also found that the scalar shrinkage estimator can lead to appreciable bias.

Although the methods we discussed have been developed for a two-dimensional interaction structure (i.e., the genetic and interaction effects are assumed to be invariant with time), they can be easily modified to allow for time-dependent effects. To allow for temporal changes in the main effects and interaction effects, one may use spline functions. For example, the mean function for Tukey's model at time (or age) of measurement t can be expressed as

$$f(\boldsymbol{\eta}(t), \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}(t), \theta(t), \boldsymbol{x}_{kt}) = \beta_0(t) + \beta^G(t)g_k + \beta^E(t)e_{kt} + \theta(t)\beta^G(t)\beta^E(t)g_k e_{kt},$$

where the genotype g_k and the exposure variable e_{kt} for subject k at time t can be treated as continuous, $\beta_0(t)$ is the baseline function, $\beta^G(t)$ and $\beta^E(t)$ are the time-varying genetic and exposure function, and $\theta(t)$ is the time-varying interaction function. These functions can be approximated by a linear combination of basis functions [41]. We plan to address the issues of estimation and testing for the temporal dynamic changes in interaction effects using alternative models in future studies.

We have proposed a new approach in the area of longitudinal GEI cohort studies. The Tukey's one-df test for non-additivity can be very powerful in terms of detecting GEI for studies where the search for GEI is based on the presence of genetic main effects (e.g., MESA), but the test can suffer from misspecification of interaction structure. The proposed shrinkage estimation procedure, on the other hand, is useful for pathway-driven GEI studies (e.g., NAS) where there is no prior knowledge of the existence of genetic main effects. It also performs well across many scenarios. Despite the advantage of efficiency, the adaptive shrinkage estimation approach still uses the same df for interaction parameters as a saturated model. As such, the increase in power by shrinking parameter estimates toward Tukey's model estimates may be limited. However, the robust performance across multiple loci with different interaction structures remain an appealing feature of such adaptive screening tests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Kraft P, Yen Y, Stram D, Morrison J, Gauderman W. Exploiting gene-environment interaction to detect genetic associations. Human Heredity. 2007; 63(2):111–119. [PubMed: 17283440]

2. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics. 2008; 64(3):685–694. [PubMed: 18162111]

3. Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. American Journal of Epidemiology. 2012; 175(3):177–190. [PubMed: 22199027]

4. Wang Y, Huang C, Fang Y, Yang Q, Li R. Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2012; 61(1):1–24.

5. Fan R, Zhang Y, Albert PS, Liu A, Wang Y, Xiong M. Longitudinal association analysis of quantitative traits. Genetic Epidemiology. 2012; 36(8):856–869. [PubMed: 22965819]

6. Zhang H. Multivariate adaptive splines for analysis of longitudinal data. Journal of Computational and Graphical Statistics. 1997; 6(1):74–91.

7. Zhang H. Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. Statistical Methods in Medical Research. 2004; 13(1):63–82. [PubMed: 14746441]

8. Zhu, W.; Cho, K.; Chen, X.; Zhang, M.; Wang, M.; Zhang, H. BMC Proceedings. Vol. 3. BioMed Central Ltd; 2009. A genome-wide association analysis of Framingham Heart Study longitudinal data using multivariate adaptive splines; p. S119

9. Xu S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics. 2007; 63(2):513–521. [PubMed: 17688503]

10. Malzahn D, Schillert A, Müller M, Bickeböller H. The longitudinal nonparametric test as a new tool to explore gene-gene and gene-time effects in cohorts. Genetic Epidemiology. 2010; 34(5):469–478. [PubMed: 20568282]

11. Mukherjee B, Ko Y, VanderWeele T, Roy A, Park S, Chen J. Principal interactions analysis for repeated measures data: application to gene–gene and gene–environment interactions. Statistics in Medicine. 2012; 31(22):2531–2551. [PubMed: 22415818]

12. Ko Y, Chudhuri P, Park S, Vokonas P, Mukherjee B. Novel likelihood ratio tests for screening gene-gene and geneenvironment interactions with unbalanced repeated-measures data. Genetic Epidemiology. 2013; 37:581–591. [PubMed: 23798480]

13. 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. Human Genomics. 2010; 4(5):302–326. [PubMed: 20650819]

14. Tukey J. One degree of freedom for non-additivity. Biometrics. 1949; 5(3):232–242.

15. Maity A, Carroll R, Mammen E, Chatterjee N. Testing in semiparametric models with interaction, with applications to gene–environment interactions. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009; 71(1):75–96.

16. Kooperberg C, LeBlanc M. Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genetic Epidemiology. 2008; 32(3):255–263. [PubMed: 18200600]

17. Murcray C, Lewinger J, Gauderman W. Gene-environment interaction in genome-wide association studies. American Journal of Epidemiology. 2009; 169(2):219–226. [PubMed: 19022827]

18. 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. The American Journal of Human Genetics. 2006; 79(6):1002–1016.

19. Barhdadi A, Dubé M. Testing for gene-gene interaction with AMMI models. Statistical Applications in Genetics and Molecular Biology. 2010; 9(Article 2):1–27.

20. Chen Y, Chatterjee N, Carroll R. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. Journal of the American Statistical Association. 2009; 104(485):220–233. [PubMed: 19430598]

21. Bates, D.; Watts, D. Nonlinear Regression Analysis and Its Applications. Wiley; New York: 1988.

22. Lindstrom M, Bates D. Nonlinear mixed effects models for repeated measures data. Biometrics. 1990; 46:673–687. [PubMed: 2242409]

23. Vonesh E, Carter R. Mixed-effects nonlinear regression for unbalanced repeated measures. Biometrics. 1992; 48(1):1– 17. [PubMed: 1581479]

24. Crainiceanu C, Ruppert D. Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. Journal of Multivariate Analysis. 2004; 91(1):35–52.

25. Gallant, A. Nonlinear Statistical Models. Wiley; New York: 2009.

26. Vonesh E, Wang H, Majumdar D. Generalized least squares, Taylor series linearization and fisher's scoring in multivariate nonlinear regression. Journal of the American Statistical Association. 2001; 96(453):282–291.

27. Gollob H. A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika. 1968; 33(1):73–115. [PubMed: 5239571]

28. Mandel J. A new analysis of variance model for non-additive data. Technometrics. 1971; 13(1):1–18.

29. Bell B, Rose C, Damon A. The veterans administration longitudinal study of healthy aging. The Gerontologist. 1966; 6(4):179–184. [PubMed: 5342911]

30. Franklin SS, Khan SA, Wong ND, Larson MG, Levy D. The Framingham Heart Study. Is pulse pressure useful in predicting risk for coronary heart disease? Circulation. 1999; 100(4):354–360. [PubMed: 10421594]

31. Kwong WT, Friello P, Semba RD. Interactions between iron deficiency and lead poisoning: epidemiology and pathogenesis. Science of The Total Environment. 2004; 330(1):21–37. [PubMed: 15325155]

32. Bradman A, Eskenazi B, Sutton P, Athanasoulis M, Goldman LR. Iron deficiency associated with higher blood lead in children living in contaminated environments. Environmental Health Perspectives. 2001; 109(10):1079–1084. [PubMed: 11675273]

33. Perlstein T, Weuve J, Schwartz J, Sparrow D, Wright R, Litonjua A, Nie H, Hu H. Cumulative community-level lead exposure and pulse pressure: the Normative Aging Study. Environmental Health Perspectives. 2007; 115(12):1696. [PubMed: 18087585]

34. Zhang A, Park S, Wright R, Weisskopf M, Mukherjee B, Nie H, Sparrow D, Hu H. *HFE H63D* polymorphism as a modifier of the effect of cumulative lead exposure on pulse pressure: the Normative Aging Study. Environmental Health Perspectives. 2010; 118(9):1261–1266. [PubMed: 20478760]

35. Knutson M, Wessling-Resnick M. Iron metabolism in the reticuloendothelial system. Critical Reviews in Biochemistry and Molecular Biology. 2003; 38(1):61–88. [PubMed: 12641343]

36. Chung J, Wessling-Resnick M. Molecular mechanisms and regulation of iron transport. Critical Reviews in Clinical Laboratory Sciences. 2003; 40(2):151–182. [PubMed: 12755454]

37. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic Epidemiology. 2008; 32(4):361–369. [PubMed: 18271029]

38. Bild DE, Bluemke DA, Burke GL, Detrano R, Roux AVD, Folsom AR, Greenland P, Jacobs DR Jr, Kronmal R, Liu K, et al. Multi-ethnic study of atherosclerosis: objectives and design. American Journal of Epidemiology. 2002; 156(9):871–881. [PubMed: 12397006]

39. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Mägi R, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics. 2010; 42(11):937–948. [PubMed: 20935630]

40. Fisher, R. Statistical Methods for Research Workers. Oliver & Boyd; Edinburgh: 1925.

41. Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998; 85(4):809–822.

Table 1

Power for detecting GEI and Type I error rates using Tukey's model, the proposed adaptive shrinkage estimator, and saturated interaction model under different interaction structures in 3×3 table settings (N=1200). Data were simulated under an autoregressive-1 (AR-1) correlation structure while analysis was performed under correctly specified and misspecified correlation structures (see Section 6.1 for simulation details).

Test Model	$\sigma^2 = 4$			$\sigma^2 = 8$		
True Model	Tukey LRT	Shrinkage Wald	Saturated LRT	Tukey LRT	Shrinkage Wald	Saturated LRT
Correctly Specified Correlation Structure (AR-1)						
Tukey's one-df	0.758	0.686	0.540	0.479	0.409	0.273
AMMI1	0.303	0.720	0.817	0.211	0.398	0.514
Saturated	0.094	0.690	0.806	0.086	0.325	0.459
$H_0 : \theta = 0$ (Additive)	0.047	0.042	0.053	0.053	0.043	0.051
$H_0 : \theta = 0$ (Null)	0.104	0.081	0.049	0.107	0.087	0.052
Misspecified Correlation Structure (Compound Symmetric)						
Tukey's one-df	0.730	0.640	0.496	0.435	0.376	0.244
AMMI1	0.287	0.708	0.792	0.185	0.372	0.489
Saturated	0.060	0.646	0.775	0.065	0.308	0.432
$H_0 : \theta = 0$ (Additive)	0.048	0.046	0.053	0.070	0.048	0.060
$H_0 : \theta = 0$ (Null)	0.118	0.064	0.053	0.143	0.065	0.061

Table 2

Average performance of tests using Tukey's model, saturated interaction model, and the adaptive shrinkage estimator for detecting GEI across 100 simulated SNPs under scenarios (A): all simulated GEI are of Tukey's form, (B): 2/3 of simulated GEI are of Tukey's form and 1/3 are of saturated form, and (C): 2/3 of simulated GEI are of saturated form and 1/3 are of Tukey's form

Measure	Scenario		Tukey LRT	Shrinkage Wald	Saturated LRT
Marginal Models					
True Positive Rate	(A)		0.7221	0.5766	0.4302
	(B)		0.5611	0.6317	0.5769
	(C)		0.4923	0.6699	0.7357
False Positive Rate			0.0024	0.0007	0.0006
Multivariate Models					
True Positive Rate	(A)		0.3264	0.2810	0.0706
	(B)		0.2882	0.2602	0.2247
	(C)		0.2073	0.2507	0.2911
False Positive Rate			0.0045	0.0027	0.0006

Table 3

The *p*-values of GEI tests for the top three (ranks in parentheses) single-nucleotide polymorphisms (SNPs) by using Tukey's model, the proposed shrinkage estimator, and saturated interaction model within iron gene regions in the NAS data (adjusted $a = 5.6 \times 10^{-4}$).

Bone Lead	SNP ID	Gene	Tukey LRT	Shrinkage Wald	Saturated LRT
Tibia	rs1799945	HFE	0.003 (1)	1×10^{-4} (1)	0.006 (1)
	rs2285228	DMT1	0.005 (2)	0.001 (2)	0.017 (2)
	rs3821716	MFI2	0.014 (3)	0.012	0.120
	rs422982	DMT1	0.016	0.003 (3)	0.072 (3)
Patella	rs7165535	B2M	0.001 (1)	4×10^{-4} (2)	0.014 (1)
	rs17484524	IREB2	0.002 (2)	3×10^{-4} (1)	0.021 (2)
	rs7866419	ACO1	0.009 (3)	0.005	0.054
	rs1358024	TF	0.016	0.004 (3)	0.038
	rs2304704	SLC40A1	0.044	0.030	0.034 (3)

Table 4

Findings of GEI with significant meta-analysis *p*-values for the single-nucleotide polymorphisms (SNPs) that have demonstrated significant and replicated evidence of marginal association with BMI in the MESA data (adjusted $\alpha = 1.9 \times 10^{-3}$).

Exposure	SNP ID	Gene	Tukey LRT	Shrinkage Wald	Saturated LRT
Energy Intake	rs543874	SEC16B	$<1.0 \times 10^{-8}$	$<1.0 \times 10^{-8}$	1.8×10^{-4}
	rs1558902	FTO	4.8×10^{-5}	7.4×10^{-4}	0.130
	rs10767664	BDNF	1.2×10^{-3}	0.103	0.124
Exercise	rs206936	NUDT3, HMGA1	1.4×10^{-4}	0.006	0.005