Research Article

Statistics in Medicine

Received 6 February 2014, Accepted 8 December 2014 Published online 29 December 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6401

Latent variable models for gene–environment interactions in longitudinal studies with multiple correlated exposures

Yebin Tao,*† Brisa N. Sánchez and Bhramar Mukherjee

Many existing cohort studies designed to investigate health effects of environmental exposures also collect data on genetic markers. The Early Life Exposures in Mexico to Environmental Toxicants project, for instance, has been genotyping single nucleotide polymorphisms on candidate genes involved in mental and nutrient metabolism and also in potentially shared metabolic pathways with the environmental exposures. Given the longitudinal nature of these cohort studies, rich exposure and outcome data are available to address novel questions regarding gene–environment interaction ($G\times E$). Latent variable (LV) models have been effectively used for dimension reduction, helping with multiple testing and multicollinearity issues in the presence of correlated multivariate exposures and outcomes. In this paper, we first propose a modeling strategy, based on LV models, to examine the association between repeated outcome measures (e.g., child weight) and a set of correlated exposure biomarkers (e.g., prenatal lead exposure). We then construct novel tests for $G \times E$ effects within the LV framework to examine effect modification of outcome–exposure association by genetic factors (e.g., the hemochromatosis gene). We consider two scenarios: one allowing dependence of the LV models on genes and the other assuming independence between the LV models and genes. We combine the two sets of estimates by shrinkage estimation to trade off bias and efficiency in a data-adaptive way. Using simulations, we evaluate the properties of the shrinkage estimates, and in particular, we demonstrate the need for this data-adaptive shrinkage given repeated outcome measures, exposure measures possibly repeated and time-varying gene–environment association. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: gene–environment dependence; gene–environment interaction; growth curves; latent variable model; shrinkage estimation

1. Introduction

Most common human diseases have a multifactorial etiology involving genetic factors (G) and environmental exposures (E). In recent years, many environmental cohort studies initially designed to study environmental health effects have begun to collect genetic information on study participants. One of the initial goals of the Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) project, for example, was to assess the impact of lead exposure on children's mental development. However, the ELEMENT project, now of more than 18 years duration, has expanded to include longitudinal outcomes such as anthropometry, adolescent behavior, sexual maturation, and cardiovascular health among youth [1–3]. Given the solid grounding in environmental health, measures of multiple toxicants, particularly lead, are available, and some have been measured repeatedly over time. With the lowering cost of genotyping technologies, the study has begun to genotype stored biological samples for single nucleotide polymorphisms (SNPs) along genes known to be involved in mental or nutrient metabolism [4–8]. It is now possible to interrogate the available ELEMENT data to help elucidate questions regarding how genetic makeup may exacerbate or reduce exposure effects previously identified, that is, examine gene–environment interaction ($G\times E$) [9–12]. Meanwhile, given known challenges to conduct $G\times E$ studies primarily due to sample size limitations [13], there is an increasing need for methods and modeling strategies that can exploit the complex data structure in an efficient and simultaneously robust way.

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
**Correspondence to: Yebin Tao, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.*
†E-mail: yebintao@umich.edu

In this manuscript, we develop modeling strategies to examine the joint impact of genes and multiple exposure measures on health outcomes measured repeatedly over time. To motivate and illustrate the ideas, we focus on four biomarkers of prenatal lead exposure (maternal bone lead concentrations at two sites and maternal and umbilical cord blood lead concentrations), two SNPs on the hemochromatosis (HFE) gene, and weight measured approximately every 6 months from birth to age of 4. Given that weight is measured at multiple time points approximately balanced across participants, one possible simple analysis is to look at each pair of (G, E) at a given time point by running several cross-sectional analyses using multiple linear regression. This naïve method lacks power due to the cost of multiple testing. One could alternatively use standard mixed models or generalized estimating equations to account for the complete longitudinal trajectory of outcome in one model, but again look at each pair of (G, E) separately. With a SNPs and b exposures, this will lead to $a \times b$ different models. By reducing the dimension of G, E, and Y, one can decrease the number of tests and models to be fit. With a correlated set of exposure biomarkers, measures of lead in our example, the use of latent variables (LVs) is a natural way to reduce the dimension of the exposure space. Very few papers in the literature contain examples of a multi-G-multi-E analysis in a joint multivariate-Y model.

Longitudinal studies offer more precise characterization of cumulative lifetime exposure and within-person variability than a cross-sectional or case-control study [14, 15]. However, they require careful analytic considerations as the interplay of genes and environment may change dynamically over time. That is, the effect modification role of genetic factors may become increasingly important, or their initial protective role may fade over time. In spite of the vast literature on statistical methods for analyzing $G \times E$ effects in case-control studies [16, 17], efficient alternative modeling strategies for $G \times E$ effects in longitudinal studies have received very little attention [18].

Latent variable models have been widely used in modeling longitudinal and multi-level data [19, 20]. They have also been applied in environmental health studies to characterize the health effects of a set of environmental exposures that are highly correlated, thus avoiding the multiple testing issue when dealing with each exposure individually [21, 22]. These models have great potential in testing $G \times E$ effects in the context of multiple genes and multiple correlated environmental exposures.

Currently, only a few studies have attempted LV models for studying $G \times E$ effects in cross-sectional or cohort studies [23–26]. As a recent contribution to this area, Sánchez *et al*. [26] investigated $G \times E$ effects for univariate outcome with multiple correlated exposures in a cross-sectional study. They built LV models for the exposures to deal with the exposure measurement errors and boost efficiency for testing $G \times E$ effects by reducing the number of tests and combining information across the available biomarkers. One important contribution of their paper was to allow for a separate gene–environment (G-E) dependence model that may help understand how exposures are related to genes (especially when the genes are chosen based on the metabolic pathway for the environmental exposures). To improve efficiency and protect against bias, the authors used shrinkage estimation for various specifications of the G-E association model [27–29]. Their proposed approach yielded estimates that balanced between bias and variance and provided an automated way to avoid model selection issues because a robust adaptive estimator was recommended as the default choice.

However, a general LV framework for studying $G \times E$ effects on longitudinal or multivariate outcomes has not been proposed. In this paper, we undertake the task of integrating longitudinally measured outcome, a set of correlated exposures (potentially time-varying), and genes (time-invariant genotype data measured at SNPs), as an extension to the cross-sectional framework of Sánchez *et al*. [26]. Several new and challenging enhancements are warranted. First, we propose to use LV models for both exposures and the longitudinal outcome. We model the weight trajectories using random coefficients corresponding to parametric functions of time that are chosen a priori; the use of random coefficients naturally fits into the general LV modeling framework [20]. Figure 1 is a path diagram describing the relationships between exposures biomarkers (E), latent exposures (U), genes (G), latent outcome (B), and observed longitudinal health outcome (Y). The diagram encodes the time-independent $G \times U$ term (interaction arrow directed at the random intercept B_0), as well as the three-way $G \times U \times T$ term of interest (dashed arrow directed at the latent outcome B_k). Second, in addition to combining estimates from models that assume varying degrees of G-E dependence [26], we also consider varying degrees of dependence of the variance of the longitudinal outcome on the genetic factors, denoted as G-b heteroscedasticity. The third and main novelty is to consider time-varying $G \times E$ effects. This time-dependent interaction may be captured by a linear three-way interaction term $G \times U \times T$ or by a more complex non-linear function of time. We additionally examine how adaptive shrinkage estimation can be used to gain power for detect-

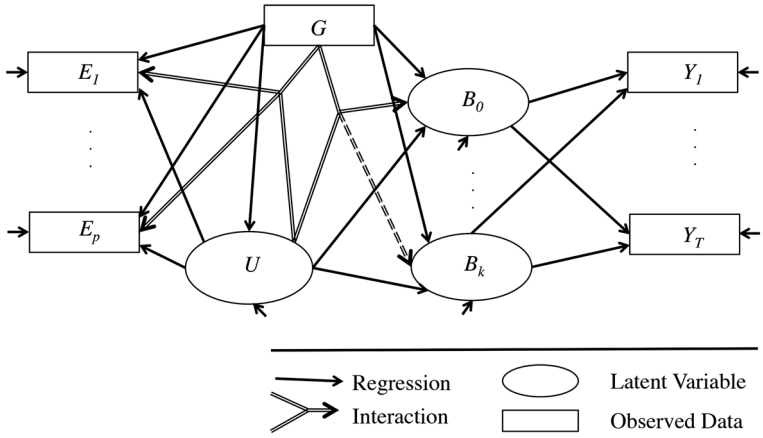

Figure 1. Path diagram showing relationships among exposure biomarkers (E), latent exposures (U), genes (G), longitudinal health outcome (Y), and latent outcome (B). The unspecified arrows indicate the possible effects of covariates and random errors.

ing this time-dependent interaction when the gene–environment association structure can vary over time. Previous work in adaptive shrinkage in longitudinal studies is limited.

We organize the rest of this paper as follows: in Section 2, we present the LV models for both a set of correlated exposures and the longitudinal outcome including gene–environment interactions. We also incorporate varying degrees of *G-E* and *G-b* heteroscedasticity into our modeling framework and present shrinkage estimation to adaptively compromise between the most parsimonious and most flexible models. In Section 3, we present a more general LV model where the exposures may vary over time. Analyses of the ELEMENT data (Section 4) and simulation studies (Section 5) bring out salient features of our modeling and $G \times E$ testing strategy. Section 6 briefly discusses the advantages and limitations of our methodology and indicates possibilities for further extensions.

2. Latent variable models for $G \times E$ studies with longitudinal outcome

2.1. Model formulation

Let Y_{ij} denote the health outcome for subject i ($i = 1, \ldots, N$) at time t_{ij} ($j = 1, \ldots, n_i$). Let U_i be an $m \times 1$ vector of latent exposures underlying the observed exposure measurements E_i ($p \times 1, p > m$). In our motivating example, $m = 1, p = 4$, and U represent prenatal lead exposure as reflected by the four lead biomarkers collected in several tissues. In other applications, variables in E may represent metabolites of a parent compound U [30], and one can propose more than one U ($m > 1$) if more than one family of exposure biomarkers are observed (e.g., mercury and polychlorinated biphenyls) [31], or if exposures are observed repeatedly over time [32]. Genetic subgroups are denoted by a categorical variable G_i. For brevity, we consider only binary G_i: for example given multiple risk alleles, this may be zero for wild type and one for at least one variant. To deal with more categories (e.g., quantiles of a polygenic risk score [33]), we can simply use multiple dummy variables. Our LV models contain two parts: the exposure model and the outcome model.

The exposure model consists of a measurement model relating the observed exposure measurements E_i to the LVs U_i

$$E_i = \nu_0 + \nu_1 G_i + \Lambda_0 U_i + \Lambda_1 G_i \cdot U_i + \delta_i \tag{1}$$

and an accompanying model with covariates V_i ($q \times 1$) that may be related to the LVs U_i

$$U_i = \alpha_0 + \alpha_1 G_i + \alpha_V V_i + \varsigma_i \tag{2}$$

Mean vectors ν and ν_1 are $p \times 1$, whereas factor loadings matrices Λ and Λ_1 are $p \times m$ (see identifiability constraints in the succeeding text). Note that (1) allows the intercepts and factor loadings to differ by genetic subgroups. The error vector δ_i has zero mean and covariance matrix Θ_G ($p \times p$) that may also depend on genetic subgroups. The hypotheses of $\Lambda_1 = 0$, $\nu_1 = 0$, and $\Theta_G = \Theta$ could be tested using

standard procedures [34, 35], which may have low power, but we avoid this model selection step by shrinkage estimation (see Section 2.2 for details). Regression coefficients α_0 and α_1 are $m \times 1$, and α_V is $m \times q$. The error vector ς_i has mean zero and covariance matrix $\Phi_G(m \times m)$. In (1) and (2), ν_1, Λ_1, and α_1 characterize the *G-E* association. We present the exposure model with coefficients independent of time, which is usually the case when we have time-invariant exposures (e.g., prenatal lead exposures). We also propose an extension of this model to time-varying exposures in Section 3.

The outcome model follows the latent growth curve modeling approach [20, 36, 37]. The longitudinal trajectory is modeled by

$$Y_{ij} = B_{0i} + \sum\nolimits_k B_{ki} f_k(t_{ij}) + \varepsilon_{ij}, \tag{3}$$

where B_{0i} is the subject-specific intercept (e.g., birth weight for subject i) and B_{ki} is the subject-specific slope corresponding to a known functional predictor term $f_k(t)$ (e.g., t, t^2 or $t^{-1}, k = X1, \ldots, K$), which captures nonlinearity in the trajectories, and between-subject variation in the growth curves. We can choose f_k's from a known set of parametric functions using model selection criteria (e.g., Bayesian information criteria) [37], prior to incorporating genes and exposures into the model. We assume that ε_{ij}'s are independently distributed with mean zero and variance σ^2 (under a balanced design, the variance can be allowed to vary over time if necessary). The latent outcomes B_{0i} and B_{ki} explain the correlation between Y_{ij}'s. We then associate B_{0i} or B_{0i} and B_{ki} with U_i and G_i conditional on a set of baseline covariates/confounders Z_i via the following models:

$$B_{0i} = \beta_{0,0} + \beta_{1,0} G_i + \beta'_{U,0} U_i + \beta'_{G\times U,0} G_i \cdot U_i + \beta'_{Z,0} Z_i + b_{0i}, \tag{4}$$

$$B_{ki} = \beta_{0,k} + \beta_{1,k} G_i + \beta'_{U,k} U_i + \beta'_{G\times U,k} G_i \cdot U_i + \beta'_{Z,k} Z_i + b_{ki}. \tag{5}$$

The random effects b_{0i} and b_{ki} have a joint normal distribution with mean zero and variance D_G, again allowed to vary across genetic subgroups. By plugging (4) and (5) into (3), one can see that parameter vectors $\beta_{G\times U,0}$ and $\beta_{G\times U,k}$ characterize the $G \times E$ effects at baseline and subsequent time points, respectively. To model the latent outcomes B_k in (5) as fixed effects, we constrain the variance of its corresponding b_k to be zero; the variance of B_0 will generally not be zero given the correlations among the repeated outcome measures. Although in the general notation of (5), we have allowed all baseline covariates Z_i to be associated with B_{ki}; in practice, this may have to be restricted by certain assumptions, either from the perspective of estimating a potentially large number of parameters $\beta_{Z,k}$ or through a priori knowledge that may rule out the effect of certain covariates on B_{ki}. Such constraints can be readily imposed by fixing subsets of $\beta_{Z,k}$ to be zero. Finally, if time-varying covariates are available, they can be added directly to (3) as fixed effects.

In summary, (1) and (2) denote the relationships among observed and latent exposures and their relationships to covariates, and (3)–(5) link observed outcomes to latent growth outcomes, and latent outcomes to latent exposures, genes, and covariates. The parameters in (1)–(5) and their meanings are listed in Table I.

Overall G × E model

Models (1)–(5) are presented as general models for $G \times E$ analysis, allowing for time-varying $G \times E$ effects. In practice, we might first explore whether there is an overall $G \times E$ effect marginalized over time, that is, to test H_0: $\beta_{G\times U,0} = 0$ in (4) while forcing $\beta_{G\times U,k}$ in (5) to be zero. We refer to this model as the 'overall $G \times E$ model'. In reference to Figure 1, the overall $G \times E$ model describes the structural relationship between G, U, and B_0 while removing the dashed line representing $G \times E \times T$ effects.

G × E × T model

To examine whether the $G \times E$ effects vary over time, we can test H_0: $\beta_{G\times U,k} = 0$ for all $k = 1, \ldots, K$ in (5). We call the model that estimates $\beta_{G\times U,k}$ the '$G \times E \times T$ model', which corresponds to the presence of the dashed line in Figure 1. However, because f_k's are chosen principally to model the trajectory of the health outcome, they may not fully capture the true functional form of the temporal variation in $G \times E$ effects. To gain a better idea of this functional form, we may resort to prior literature or exploratory methods. For example, if the outcome is measured at several fixed time points (e.g., yearly weight) and sample size is

Table I. Summary of parameters in models (1) to (5).

Model/parameter*	Interpretation
Model (1)	Measurement model relating observed exposure measurements and latent exposures
$\nu_0\,(p \times 1)$	Intercepts of exposure measurements for the genetic reference subgroup
$\boldsymbol{\nu_1}\,(p \times 1)$	Difference in intercepts of exposure measurements between genetic subgroups
$\Lambda_0\,(p \times m)$	Factor loadings for the genetic reference subgroup
$\boldsymbol{\Lambda_1}\,(p \times m)$	Difference in factor loadings between genetic subgroups
$\boldsymbol{\Theta_G}\,(p \times p)$	Covariance matrix for error vector δ, dependent on genetic subgroup G
Model (2)	Model for latent exposure given covariates
$\alpha_0\,(m \times 1)$	Intercepts of latent exposure for the genetic reference subgroup
$\boldsymbol{\alpha_1}\,(m \times 1)$	Difference in intercepts of latent exposure between genetic subgroups
$\alpha_V\,(m \times q)$	Association between latent exposure and covariates V
$\boldsymbol{\Phi_G}(m \times m)$	Covariance matrix for error vector ς, dependent on genetic subgroup G
Model (3)	Model for observed outcome measurements given latent outcomes
B_{0i}	Subject-specific intercept for the health trajectory
B_{ki}	Subject-specific slope for the health trajectory
σ^2	Variance of residual error ε
Model (4)/(5)	Structural model linking latent exposures to latent outcomes
$\beta_{0,\bullet}$	Mean intercept/slope for the health trajectory in the genetic reference subgroup
$\boldsymbol{\beta_{1,\bullet}}$	Difference in mean intercept/slope for the health trajectory between genetic subgroups
$\beta_{U,\bullet}$	Association between latent outcomes and latent exposures for the genetic reference subgroup
$\boldsymbol{\beta_{G\times U,\bullet}}$	Difference between genetic groups in the association between latent outcomes and latent exposures
$\beta_{Z,\bullet}$	Association between latent outcomes and covariates Z
$\mathbf{D_G}$ $(K+1)\times(K+1)$	Covariance matrix for random effects b_0 and b_k, dependent on genetic subgroup G

*Parameters charactering dependence on genetic subgroups are in bold; vectors and matrices have dimension listed.

adequate, we may model the $G \times E$ effect on the outcome at each distinct time point separately (i.e., time treated as dummy variables). We then plot the $G \times E$ coefficients against time in a meta-regression analysis to gauge a suitable functional form of time-varying interaction. We may need to include a function of time, not necessarily a subset of the f_k's, to be incorporated into (5) to better model the $G \times E \times T$ effect. We use our data example to illustrate such exploratory strategies.

Identifiability

To make the LV models identifiable, we put standard constraints on the model parameters [22, 38]. For instance, typical constraints involve having factor loading matrices be prespecified as block diagonal, which means that a given observed exposure reflects only one LV, letting the first element of each nonzero block be one, and fixing the first element of the intercepts to be zero. These constraints fix the mean and scale of the LV, which are otherwise not identifiable. We apply these types of constraints to the measurement model (1). Specifically, we constrain the first entries of each block of Λ_0 to be one and the corresponding entries of ν_0 to zero so that the mean and scale of the LVs in the reference genetic subgroup are identifiable. However, as G appears also in (2), more constraints are needed. We additionally fix the first entries of the corresponding blocks of Λ_1 to be zero such that the units of the LV are the same in both groups; this ensures identifiability of the variance of the LV among the genetic reference subgroup,

$\Phi_{G=0}$ in (2). Similarly, we constrain the corresponding entries of ν_1 to be zero such that the difference in means of the latent exposures between the genetic subgroups (i.e., α_1 in (2)) is identifiable. For instance, in the context of prenatal lead exposure (Section 4), the factor loading corresponding to patella lead in Λ_0 is set to 1 and the difference in the factor loading for patella lead between the genetic subgroups, element in Λ_1, corresponding to patella lead is zero so that the units of the LV are in the units of patella lead concentration (μgPb/g) in both groups. We also assume that the off-diagonal elements of Θ_G are zero, reflecting conditional independence between E_i's given U_i. Note that this conditional independence is not strictly required and can be relaxed given strong a priori knowledge (e.g., use of the same laboratory or other circumstances that may give rise to additional correlation among E_i's). If sample size is small, we can consider additional constraints to reduce the number of parameters to be estimated, for example, forcing the elements of Θ_G to be identical across genetic subgroups, that is, $\Theta_{G=0} = \Theta_{G=1}$.

2.2. Likelihood and estimation

Let $O_i = (Y_i^{'}, E_i^{'})^{'}$ and θ be the vector of all model parameters. Assuming normality for all residuals, and integrating over the LV, the joint marginal distribution of the observed outcomes and exposures for subject i has a multivariate normal density $f(O_i|G_i, t_i, V_i, Z_i; \theta)$ (see Supplementary Materials, Likelihood, for details). The log likelihood of θ is then, $l(\theta) = \sum_{i=1}^{N} \log f(O_i|G_i, t_i, V_i, Z_i; \theta)$. We apply the standard maximum likelihood estimation (MLE) to our LV models, using the R package *lavaan* by Rosseel (2012) (http://lavaan.org) [39]. The parameter estimates are obtained by maximizing $l(\theta)$. Variances for parameters can be obtained by inverting the information matrix $I(\theta) = -E(\partial^2 l(\theta)/\partial\theta\partial\theta^{'})$. The codes for implementing the methods are available at http://www-personal.umich.edu/∼brisa/.

G-E dependence and G-b heteroscedasticity

Our LV models for exposures and longitudinal outcome are presented in a general way, allowing for full dependence on the genetic factors. To reduce the dimensionality of the parameter space, we may impose a certain degree of *G-E* independence as well as *G-b* homoscedasticity. These assumptions may hold for external exposures (air pollution and heavy metals) and a set of genes unrelated to the exposure but may not be so plausible for behavioral exposures/outcomes and genes in the same metabolic pathway. The assumption of *G-E* independence and *G-b* homoscedasticity can potentially boost the power of tests for interaction. However, the estimates of interaction may be seriously biased when the underlying assumption is violated [27, 40]. In our study, the mean and variance of the subject's birth weight and rate of weight gain, as well as the prenatal lead exposure, may not all be independent of genetic factors. The HFE gene shares a common metabolic pathway with lead exposure [41] and may potentially induce higher variance in the outcomes. The *G-b* heteroscedasticity has not been previously discussed in the literature, but we consider it possible, because groups defined by G could influence not only the outcome mean but also its variance, and furthermore, misspecifying outcome variance when latent predictors are in the model could lead to substantial bias in the regression coefficients [32].

It is difficult to determine the plausibility of the *G-E* dependence and *G-b* heteroscedasticity based on current data. A convenient and automated way is to model varying degrees of dependence and use shrinkage to obtain an estimator that balances bias and efficiency. In our study, we will first build the model under the most restrictive assumption that all the parameters are homogeneous across genotypes. We use 'AI' to denote the assumption of *G-E* independence and *G-b* homoscedasticity

$$\text{AI}: (\nu_1, \Lambda_1, \Theta_G, \alpha_1, \Phi_G, \text{D}_G) = (0, 0, \Theta, 0, \Phi, \text{D})$$

In the second step, we relax all the constraints and build the model under the assumption of complete *G-E* dependence and *G-b* heteroscedasticity, denoted as 'AD'. We consider the possibility that the variances of the latent outcomes also depend on gene, that is, all parameters may vary across genetic subgroups.

Clearly, the AD model has many more parameters than the AI model. The AI model may improve efficiency for interaction estimates but could introduce bias if the assumed independence constraints are incorrect. On the other hand, the problem with the AD model is that larger sample sizes are needed to have precise estimates of the parameters of interest. If the sample size is modest, we may put constraints on parameters like Θ_G. After estimating AI and AD, it is not straightforward to assess relative model fits for these two models. For instance, a simple two-stage approach of first testing the plausibility of parameter constraints using current data and then proceeding with AI or AD will incur a high Type-I error

rate, because the study is likely underpowered to detect significant differences across genetic subgroups for all parameters. Adaptive shrinkage of the robust AD estimators to efficient AI estimators appears to be an attractive alternative.

Shrinkage estimation

We use shrinkage estimation to combine MLE estimates under assumptions of AI and AD following Chen *et al*. [29]. Denoting the two estimators by $\hat{\theta}_{AI}$and $\hat{\theta}_{AD}$, we define

$$\hat{\theta}_{SK} = \hat{\theta}_{AD} + K^{MV}\left(\hat{\theta}_{AI} - \hat{\theta}_{AD}\right) \tag{6}$$

as the shrinkage estimator (SK). The multivariate shrinkage (MV) weights are $K^{MV} = \hat{V}(\hat{V}+\hat{\psi}\hat{\psi}^T)^{-1}$ with $\hat{\psi} = \hat{\theta}_{AI} - \hat{\theta}_{AD}$. $\hat{V}$ is the estimated asymptotic variance matrix of the estimated difference $\hat{\psi}$. Note that (6) is only defined for common parameters in the two models, and K^{MV} is an $s \times s$ matrix (s is the dimension of $\hat{\theta}_{AI}$). An alternative way to obtain shrinkage weights is to calculate them separately for each parameter so that we only need to deal with parameters of primary interest. This is called 'component-wise (CW)' shrinkage, whose weights (K^{CW}) depend only on the variance and bias related to that component, that is, $K_l^{CW} = \hat{V}_l/\left(\hat{V}_l + \hat{\psi}_l^2\right)$ $(l = 1, \ldots, s)$ [29]. Because the SK depends on two correlated estimators $\hat{\theta}_{AD}$ and $\hat{\theta}_{AI}$, its variance is approximated using the multivariate delta method [26, 29]. In a given data analysis, the variance can also be straightforwardly obtained via bootstrap.

To understand how the SK works, we need to carefully examine the weights. As we know, $\hat{\theta}_{AD}$ is unbiased but may not be efficient due to the large number of parameters in the model. This is particularly a problem if the sample size is modest. $\hat{\theta}_{AI}$, on the other hand, is usually efficient but can be seriously biased if AI is violated. When the bias is large, the weight K_l^{CW} goes toward zero, and one would favor $\hat{\theta}_{AD}$. Otherwise, one will favor the more efficient $\hat{\theta}_{AI}$. This is a typical bias-variance tradeoff [28]. The CW shrinkage tends to have better efficiency than the MV shrinkage, because K^{MV} uses the full matrix $\hat{V}$ and large sampling errors in the off-diagonals of $\hat{V}$ outweighs the potential efficiency gain from MV shrinkage [29]. Sánchez *et al*. [26] has shown that in the cross-sectional setting, the MV shrinkage has larger mean squared error (MSE) despite smaller bias. Given the longitudinal setting in our study that leads to a much larger number of parameters and therefore larger uncertainty in the off-diagonal elements of $\hat{V}$, we use CW shrinkage in all our subsequent development.

3. Extension to time-varying exposures

In Section 2, we have considered only time-independent exposures. Here, we present the extension of our method to time-varying exposures while all other notations remain the same. For subject i at time t_{ij}, let E_{ij} denote a $p \times 1$ vector of observed exposure measurements and U_{ij} be an $m \times 1$ vector of latent exposures. Then, models (1) and (2) can be rewritten as

$$E_{ij} = \nu_0 + \nu_{1j}G_i + \Lambda_0 U_{ij} + \Lambda_{1j}G_i \cdot U_{ij} + \delta_{ij} \tag{1a}$$

and

$$U_{ij} = \alpha_0 + \alpha_{1j}G_i + \alpha_V V_i + \varsigma_{ij}. \tag{2a}$$

Note that the regression coefficients unrelated to G are the same as those in (1) and (2) as we assume the relationship between the observed exposure measurements and the underlying latent exposures for the reference genetic subgroup ($G = 0$) is constant over time. Meanwhile, by allowing coefficients ν_{1j}, Λ_{1j}, and α_{1j} related to G to vary over time, we incorporate the possibility of the time-varying G-E association, that is, varying effect of genetic factors on exposure biomarkers over time. With age, different exposures may be metabolized differently inducing such time-varying G-E association [18]. The error vectors δ_{ij} and ς_{ij} both have mean zero, and their variances can be either time-independent (Θ_G and Φ_G, respectively) as in (1) and (2) or time-varying ($\Theta_{G,j}$ and Φ_{Gj}, respectively) depending on the application context. However, with moderate sample size, we may have to build more parsimonious models. For example, we may constrain $\nu_{1j} = \nu_1$, $\Lambda_{1j} = \Lambda_1$, $\alpha_{1j} = \alpha_1$, $\Theta_{G,j} = \Theta_G$ or $\Phi_{Gj} = \Phi_G$. In this way, we are averaging the G-E association over time.

Given time-varying exposures, the outcome model can be modified as

$$Y_{ij} = \beta_{G,j}G_i + \beta_{U,j}U_{ij} + \beta_{G\times U,j}G_i \cdot U_{ij} + B^*_{0i} + \sum\nolimits_k B^*_{ki} f_k(t_{ij}) + \varepsilon_{ij}, \tag{3a}$$

where $B^*_{0i} = \beta_{0,0} + \beta'_{Z,0}Z_i + b_{0i}$ (4a) and $B^*_{ki} = \beta_{0,k} + \beta'_{Z,k}Z_i + b_{ki}$ (5a). Note that in contrast to (3)–(5), we separate the G and U terms from the functional terms involved in the temporal growth trajectory and allow separate regression coefficients corresponding to G and U terms to vary over time. We can also assume that $\beta_{G,j}$, $\beta_{U,j}$, and $\beta_{G\times U,j}$ follow some specific parametric function of time. For example, if we believe that there is a linear trend of $G\times E$ effect over time, we would model $\beta_{G\times U,j}$ to be a linear function of time. Notations of parameters of covariates and variance components remain unchanged.

4. Example: weight growth prenatal lead exposure and the HFE gene

We used pooled data from ELEMENT, including three sequentially enrolled longitudinal birth cohorts recruited between 1994 and 2005 at maternity hospitals serving low-income to moderate-income populations in Mexico City [1]. We focused our analysis on children's weight as the health outcome, which was measured longitudinally from birth to 48 months. Prenatal lead exposure is the main exposure of interest, which has been demonstrated to have deleterious effects on birth weight [42] and weight trajectories [1]. Prenatal lead exposure biomarkers were collected on the mother and child, including lead concentrations in umbilical cord blood and maternal blood at delivery, and maternal bone lead concentrations at two bone sites (patella and tibia) [1,42]. Among these four biomarkers, patella lead concentration is viewed as a better indicator of cumulative fetal exposure for two reasons: (i) bone lead biomarkers are more indicative of cumulative fetal exposure during pregnancy because blood lead primarily reflects the last three months of exposure due to its relatively shorter half-life and (ii) patella is preferred over tibia bone lead concentration because trabecular bone, such as the patella bone, has a faster turnover compared with cortical bone (such as tibia), and thus more closely reflects exposure to the fetus during pregnancy [43].

Two known SNPs, C282Y and H63D, on the HFE gene [41] were considered as effect modifiers of the lead-growth association. Because of sparsity of data, we applied dominant models for genetic susceptibility and created a single indicator variable: zero for wild type on both SNPs and one for at least one copy of either of the risk alleles. Although the HFE gene has been primarily linked with iron metabolism [41], it has also been shown to modify lead absorption in an age-dependent fashion [8,44]. As such, it is possible that the correlations among exposure biomarkers may differ between wild types and variants, implying that the exposure model coefficients may differ by genetic subgroup. The HFE gene has also been shown to have joint health effects with lead exposure [6,44].

To be included in this analysis, children must be genotyped, have weight measured at more than two time points, and have at least one of the four prenatal exposures biomarkers ($N = 758$). We focused our analysis on time points at birth and months 4, 12, 18, 24, 30, 36, and 48 when the cohorts had over 60% complete weight measurements. We followed Afeiche *et al.* [1] in choosing additional covariates based on biological relevance. Missing data on covariates was imputed five times [45], and parameter estimates from each imputed dataset were combined by standard formulae [46].

The procedure to conduct our analysis can be summarized into the following steps:

(1) We defined an LV for prenatal lead exposure using the four lead biomarkers (Figure 1). Because patella lead is a preferred biomarker of cumulative lead exposure during pregnancy as mentioned before, the mean and scale of the LV are set equal to those of patella lead using the identifiability constraints described in Section 2.1.
(2) We modeled the weight growth trajectories by regressing weight on time in the linear mixed model adjusting for selected covariates. We did not consider the genetic factor or lead exposures at this stage. We incorporated functions of t, t^2, and $\log(t)$ (hence, $K = 3$) in the model to capture the time trend of the mean weight trajectory, which have been typically used to model growth trajectories in children [37]. We further examined the model fit with residual plots.
(3) We fit the overall $G \times E$ model to the data to examine the marginal $G \times E$ effect averaged over the whole study period.
(4) Given the balanced design with respect to time (except for missed visits), we explored the $G \times E$ effect for each time point by treating time as dummy variables, denoted as the 'discrete time $G \times E$

model'. Specifically, for subject i with genetic category G_i at time j ($i = 1, \ldots, 758$ and $j = 1, \ldots, 8$ in our example),

$$Y_{ij} = \gamma_{G,j} G_i + \gamma_{U,j} U_i + \gamma_{G\times U,j} G_i \cdot U_i + \gamma_j' X_{ij} + e_{ij} \tag{7}$$

where $\gamma_{G,j}, \gamma_{U,j}$ and $\gamma_{G\times U,j}$ are regression coefficients for genetic factor G_i latent exposure U_i and the interaction term, respectively, and X_{ij} and γ_j are vectors of covariates (including intercept) and regression coefficients, respectively. The error vector $e_i' = (e_{i1}, \ldots, e_{i8}) \sim N(0, \Sigma_{8\times 8})$, where $\Sigma_{8\times 8}$ was unstructured in our analysis, based on model fit. By plotting $\gamma_{G\times U,j}$ against t_j, we found that although the outcome model included multiple $f_k(t)$'s (i.e., t, t^2 and $log(t)$), the $G\times U\times T$ term could be described with a simple linear function of time $h(t) = t$, a subset of $f_k(t)$'s. In other applications, it is possible that $h(t)$ is completely different from $f_k(t)$'s.

(5) We fit the $G\times E\times T$ model to the data by forcing the coefficients of $G\times U$ terms corresponding to t^2 and $log(t)$ to be zero while retaining all other terms, including the $G\times U$ term for t, in (5).

Results from the ELEMENT analysis

The weight data at birth is complete while at other ages, the percent of missingness ranges from 12.5% to 36.7% (Supplemental Materials, Table S1). The weight increases over time but with decreasing rate, and the variance is increasing (Figure 2). By applying the overall $G\times E$ model, we do not find a significant overall $G\times E$ effect (Table II). However, there is a decreasing trend in $G\times E$ effect over time when fitting the discrete time $G\times E$ model and plotting $\hat{\beta}_{G\times U}$ against time (aforementioned step 3, Figure 3). The negative $\hat{\beta}_{G\times U}$ indicates that carrying a variant for the HFE gene may exacerbate the deleterious effects of prenatal lead exposures on weight growth, particularly after one year from birth. The decreasing trend indicates that the degree of exacerbation may escalate over time. Comparing the results from the overall $G\times E$ model and discrete time $G\times E$ model, it is clear that if we ignore the time-varying component in the $G\times E$ effect, we would completely miss the $G\times E$ signal.

Guided by the discrete time $G\times E$ model, we fitted the $G\times E\times T$ model with linear T (Table II). We found a statistically significant $\hat{\beta}_{G\times U\times T}$, indicating that the modification of lead-growth association by HFE status is age-dependent. The estimates and robust standard errors under AI are fairly similar to those under AD, likely implying weak *G-E* and *G-b* associations. The negative estimates indicate that, on average, the variants compared with wild types may suffer from even further reductions in weight gain in association with higher lead exposure (i.e., further impairment in growth). Via shrinkage estimation, the further reduction in weight gain associated with 10μgPb/g higher patella bone concentration is estimated to increase by 96.9 g every 6 months. In other words, comparing two children that only differ in HFE status, if their patella lead level were both to increase by 10μgPb/g, the one in the variant group would gain less weight than the one in the wild-type group, and their difference in weight gain would increase by 96.9 g every 6 months. Using appropriate linear combinations of $\hat{\beta}_{G\times U}$ and $\hat{\beta}_{G\times U\times T}$, we also calculated estimates for the $G\times E$ association at each age as shown in Figure 3. Compared with the discrete time $G\times E$ model, the estimates from the $G\times E\times T$ model are generally similar, while the standard errors

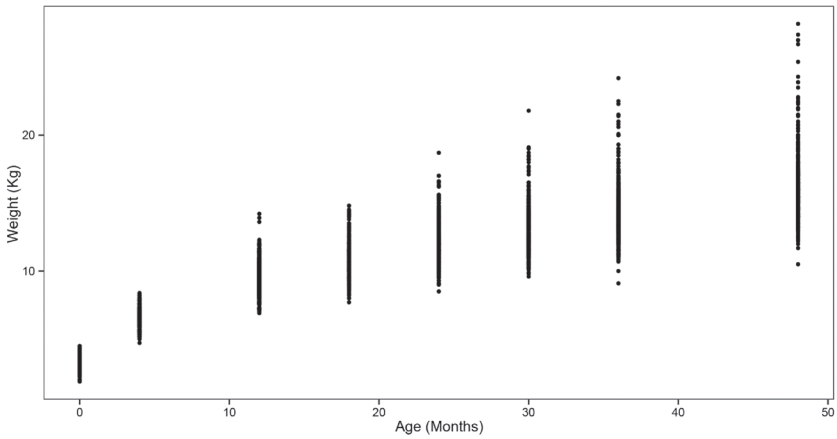

Figure 2. Scatterplot of children?s weight (kg) by age (months) in the ELEMENT example.

Table II. Parameter estimates and robust standard errors obtained using MLE derived assuming AI and AD, and shrinkage estimates (SK) for ELEMENT data.

Estimation method	Overall $G \times E$ model		$G \times E \times T$ model with linear T[a]	
	$\hat{\beta}_U$	$\hat{\beta}_{G\times U}$	$\hat{\beta}_{G\times U}$	$\hat{\beta}_{G\times U\times T}$
AI	15.2 (27.2)	−25.1 (45.1)	−20.7 (43.7)	−96.6 (43.4)*
AD	17.3 (29.6)	−21.6 (47.8)	−16.7 (46.6)	−97.1 (45.5)*
SK	15.5 (27.4)	−24.4 (45.5)	−18.0 (45.0)	−96.9 (43.9)*

Given the identifiability constrains on the exposure model, coefficients represent changes in weight (g) associated with 10μgPb/g higher patella lead level. Models are adjusted for maternal age, height, calf circumference, parity, education, marital status, lifestyle, and calcium treatment, as well as children's gestational age, cohort, and repeated measures of height.

[a]T has been rescaled to represent changes per 6 months.

*p-value < 0.05.

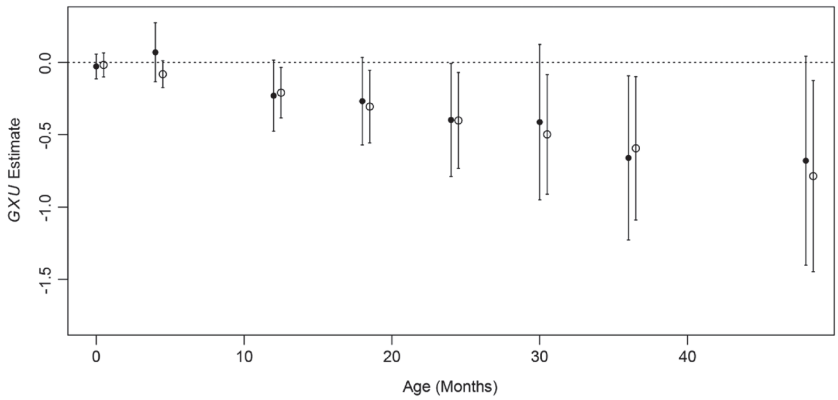

Figure 3. Shrinkage estimates and 95% confidence interval for gene–environmental interaction over time, representing the difference in the lead-weight association in HFE gene variants compared with wild types at each age. The black circles are from the analysis with the discrete time $G \times E$ model and the empty circles are from analysis with the $G \times E \times T$ model linear in T. The dashed line indicates null effect.

are much smaller, resulting in improved power. This example illustrates the value of thinking about time-dependent features in interaction patterns as opposed to an interaction term that is assumed to be constant over time.

5. Simulation studies

We carried out simulation studies to assess the bias, efficiency, MSE, power, and Type-I error properties of our methods. Specifically, we considered three settings of simulations where the true models are as follows: (i) the null model of no $G \times E$ effect at any time point; (ii) time-independent $G \times E$ effect; and (iii) time-varying $G \times E$ effect.

5.1. Simulations with time-invariant exposures

We used a sample size of $N = 500$ that is typical in environmental health studies. The genetic subgroup was generated as a binary variable with prevalence 0.3: zero for all wild types and one for having at least one variant. We simulated four exposure measurements, assuming that they measured one latent exposure. The latent exposure, under AD, was simulated from $N(0, 1)$ for the wild type group and $N(1, 2)$ for the variant group (i.e., $\alpha_1 = 1$, $\Phi = 1$, and $\Phi_1 = 2$ in (2)), while under AI, $N(0, 1)$ for both groups. The latent outcomes were generated as random intercept and random slope for each subject ($K = 1$ and $f(t) = t$ in (3)). The longitudinal outcomes were generated at five equally spaced time points ($t = 0, \ldots, 4$).

Our parameters of interest are $\beta_{G\times U}$, the baseline difference in exposure effect between genetic subgroups, and $\beta_{G\times U\times T}$, the difference in the $G \times E$ effect by unit increase of time. The true interaction parameters for the three settings were (i) $\beta_{G\times U} = \beta_{G\times U\times T} = 0$; (ii) $\beta_{G\times U} = 2$, $\beta_{G\times U\times T} = 0$; and (iii) $\beta_{G\times U} = 2$, $\beta_{G\times U\times T} = -1$ with the $G \times U \times T$ term assumed to be linear in T, respectively. For all three settings, we investigated the properties of our estimators when the underlying data generating mechanism was AI or AD. For setting 3, we further considered two scenarios: (i) data was simulated under *G-E* independence and *G-b* heteroscedasticity, denoted as 'APD1', to isolate the role of *G-b* heteroscedasticity alone in determining the operating characteristics of the methods and (ii) data were simulated under *G-E* partial dependence (only a subset of exposures interactive with genes) and *G-b* homoscedasticity, denoted as 'APD2'. Details regarding the choice of remaining parameters and how they varied across genetic subgroups can be found in the Supplemental Materials, Table S3. For settings 1 and 2, we fitted the overall $G \times E$ model, and for setting 3, we also fitted the $G \times E \times T$ model. We summarized the results based on 500 replicated datasets.

Simulation results

Table III presents the estimates from the overall $G \times E$ model for all three settings. When there is no interaction in the true model ($\beta_{G\times U} = \beta_{G\times U\times T} = 0$) and the data are generated under AI, all methods have acceptable biases and Type-I error rates. However, when the data are generated under AD, only the MLE assuming AD and the SK retain probability of rejection of a null hypothesis (P(R)) close to 0.05 at the nominal level of 0.05. As expected, MLE assuming AI has bias and inflated Type-I error in this scenario. When the overall $G \times E$ effect exists ($\beta_{G\times U} = 2$ and $\beta_{G\times U\times T} = 0$), all approaches have power of at least 0.75 except MLE assuming AI with data generated under AD. This loss of power is primarily due to severe bias in $\hat{\beta}_{G\times U}(-36.0\%)$. In the first two settings, the SK performs very similar to the correct estimator, in terms of bias, MSE, and power (P(R)). For setting 3 ($\beta_{G\times U} = 2$ and $\beta_{G\times U\times T} = -1$) when the overall $G \times E$ model is misspecified, none of these approaches have adequate performance.

Table IV presents the results from fitting the model with linear $G \times U \times T$ term to the data generated in setting 3. Under data scenario AI or APD1, all approaches have satisfactory performance with the correct approach being the most unbiased and efficient. However, under data scenario AD or APD2, MLE assuming AI is seriously biased for both $\beta_{G\times U}$ and $\beta_{G\times U\times T}$. The similarities between scenarios AI and APD1 and between AD and APD2 indicate that *G-E* dependence may be more critical than *G-b* heteroscedasticity for determining the performances of our methods. For each of the four scenarios, the SK maintains very acceptable bias and efficiency

5.2. *Simulations with time-varying exposures*

We also conducted simulations assuming exposures are measured concurrently with the outcome, as postulated in (1a)–(2a). Again, we generated outcomes and exposures at five time points. We simulated

Table III. Bias (or percent bias), MSE, and rejection probabilities (P(R)) for estimates of time-independent $G \times E$ effect ($\beta_{G\times U}$) with data simulated under AI or AD sample size $N = 500$, with 500 replicates.

		Parameter setting[b]								
		1			2			3		
Data scenario	Estimation method[a]	Bias	MSE	P(R)	Bias%	MSE	P(R)	Bias%	MSE	P(R)
AI	AI[c]	0.00	0.33	0.04	1.2	0.35	0.93	−47.5	1.24	0.42
	AD[c]	0.03	0.38	0.05	3.0	0.47	0.90	−38.3	1.08	0.37
	SK[d]	0.02	0.36	0.04	1.4	0.41	0.92	−42.5	1.14	0.41
AD	AI[c]	−0.53	0.68	0.19	−36.0	0.92	0.62	−72.0	2.20	0.19
	AD[c]	0.01	0.48	0.05	1.1	0.69	0.80	−15.1	1.85	0.52
	SK[d]	−0.11	0.47	0.08	−7.1	0.68	0.75	−23.3	1.59	0.40

[a]Estimation uses the model with only time-independent $G \times U$ term.
[b](1) $\beta_{G\times U} = \beta_{G\times U\times T} = 0$; (2) $\beta_{G\times U} = 2$, $\beta_{G\times U\times T} = 0$; and (3) $\beta_{G\times U} = 2$, $\beta_{G\times U\times T} = -1$. $f(t) = t$ for all three settings.
[c]MLE derived assuming AI or AD.
[d]Shrinkage estimator combining AI and AD.

Table IV. Percent bias, variance ratios (Var.R), MSE, and rejection probabilities (P(R)) for estimates of $G \times E$ effect with data simulated under AI, AD APD1, and APD2[a].

Data scenario	Estimation method[b]	$\hat{\beta}_{G\times U}$ Bias%	Var.R	MSE	P(R)	$\hat{\beta}_{G\times U\times T}$ Bias%	Var.R	MSE	P(R)
AI	AI[c]	0.5	1(Ref)	0.43	0.82	−2.3	1(Ref)	0.11	0.79
	AD[c]	2.4	(1.16)	0.50	0.82	−3.8	(1.04)	0.12	0.73
	SK[d]	1.2	(1.07)	0.45	0.81	−3.0	(1.02)	0.11	0.75
AD	AI[c]	−26.2	1(Ref)	0.85	0.66	−18.9	1(Ref)	0.16	0.90
	AD[c]	2.0	(1.32)	0.75	0.77	2.1	(1.03)	0.13	0.86
	SK[d]	−6.8	(1.22)	0.71	0.72	−6.6	(1.1)	0.12	0.89
APD1	AI[c]	5.5	1(Ref)	0.48	0.86	−12.3	1(Ref)	0.18	0.63
	AD[c]	2.8	(1.13)	0.53	0.80	−4.5	(1.01)	0.15	0.64
	SK[d]	3.0	(1.06)	0.50	0.82	−6.3	(1.00)	0.16	0.64
	APD1[c]	1.1	(0.97)	0.45	0.82	−4.0	(0.96)	0.14	0.69
APD2	AI[c]	−24.6	1(Ref)	0.73	0.71	−18.2	1(Ref)	0.15	0.86
	AD[c]	1.0	(1.17)	0.57	0.84	−2.7	(1.07)	0.13	0.89
	SK[d]	5.3	(1.16)	0.58	0.78	−6.7	(1.01)	0.13	0.88
	APD2[c]	0.9	(1.01)	0.51	0.85	−2.5	(0.98)	0.11	0.90

$\beta_{G\times U} = 2$, $\beta_{G\times U\times T} = -1$, $f(t) = t$, and sample size $N = 500$, with 500 replicates.
[a]APD1: *G-E* independence and *G-b* heteroscedasticity; APD2: *G-E* partial dependence and *G-b* homoscedasticity.
[b]Estimation uses the model with both linear $G \times U$ and $G \times U \times T$ terms.
[c]MLE derived assuming AI, AD, APD1 or APD2
[d]Shrinkage estimator combining AI and AD.

Table V. Bias (or percent bias), variance ratios (Var.R), MSE, and rejection probabilities (P(R)) for estimates of $G \times E$ effect using data simulated with time-varying exposures and time-varying *G-E* association.

Parameter setting	Estimation method[a]	$\hat{\beta}_{G\times U}$ Bias	Var.R	MSE	P(R)	$\hat{\beta}_{G\times U\times T}$ Bias	Var.R	MSE	P(R)
$\beta_{G\times U}$ =	True[b]	0.01	1(Ref)	0.29	0.03	0.00	1(Ref)	0.16	0.03
$\beta_{G\times U\times T}$ =	AI[c]	−0.09	(1.02)	0.31	0.05	−0.22	(0.98)	0.28	0.09
0	AD1[c]	0.03	(1.12)	0.33	0.04	0.02	(1.10)	0.21	0.04
	AD2[c]	0.11	(1.03)	0.30	0.05	−0.10	(1.03)	0.20	0.05
	SK1[d]	−0.03	(1.06)	0.31	0.04	−0.03	(1.18)	0.22	0.05
	SK2[d]	−0.02	(1.03)	0.30	0.05	−0.12	(1.01)	0.21	0.07
$\beta_{G\times U}$ = 2	True[b]	0.5%	1(Ref)	0.32	0.90	−1.6%	1(Ref)	0.21	0.69
$\beta_{G\times U\times T}$ = −	AI[c]	−9.5%	(1.01)	0.38	0.89	−32.2%	(0.98)	0.35	0.76
1	AD1[c]	−2.0%	(1.21)	0.40	0.87	1.2%	(1.16)	0.26	0.60
	AD2[c]	2.9%	(1.08)	0.35	0.92	−16.7%	(1.01)	0.28	0.72
	SK1[d]	−3.1%	(1.16)	0.39	0.87	−3.9%	(1.12)	0.28	0.67
	SK2[d]	0.9%	(1.06)	0.34	0.90	−20.1%	(1.01)	0.30	0.73

Sample size $N = 500$ with 500 replicates.
[a]Estimation uses the model with both linear $G \times U$ and $G \times U \times T$ terms.
[b]Model with true *G-E* association: AI for time 1 and 2, and AD for time 3, 4, and 5.
[c]MLE derived assuming AI, AD with time-varying *G-E* association (AD1), or AD with time-independent *G-E* association (AD2).
[d]Shrinkage estimator combining AI and AD1 (SK1), or AI and AD2 (SK2).

exposures for each time point separately and also incorporated time-varying *G-E* association in the true model: AI for time points 1 and 2 and AD for time points 3, 4, and 5 (details on parameter setting can be found in the Supplemental Materials, Table S3). We set *G-b* heteroscedasticity for all time points. We considered two situations where there was no $G \times E$ effect ($\beta_{G\times U} = \beta_{G\times U\times T} = 0$) or there was time-varying $G \times E$ effect ($\beta_{G\times U} = 2$ and $\beta_{G\times U\times T} = -1$). Then, we applied two types of AD models: one with time-varying *G-E* association (AD1) and the other with time-independent *G-E* association (AD2). We

also calculated MLE assuming AI and used shrinkage estimation to combine AD1 with AI, denoted as SK1, as well as AD2 with AI, denoted as SK2. As a comparison benchmark, we analyzed the data with the correct *G-E* and *G-b* association (only possible in a simulation study where we know the truth).

Simulation results

Under the setting with no $G \times E$ effect, we find that all the methods have P(R) (Type-I error) close to 0.05 at the significance level of 0.05 (Table V). However, MLE assuming AI, as a parsimonious method with the most restricted set of assumptions, leads to large bias for $\beta_{G\times U\times T}$. MLE assuming AD2 also has some bias due to misspecification of the *G-E* association. When there is time-varying $G \times E$ effect, we find different results for $\beta_{G\times U}$ and $\beta_{G\times U\times T}$. Despite some bias from MLE assuming AI, all other methods show satisfactory performance for $\beta_{G\times U}$. In contrast, the power for $\beta_{G\times U\times T}$ is lower (0.60–0.76). MLEs assuming AI and AD2 yield biased estimates, while MLE assuming AD1 has much less bias but also reduced power, likely due to the large number of parameters (124 for AD1 versus 83 for correct model). SK1 has relatively acceptable performance in bias and power, when compared with SK2 and the correct method, which indicates that when dealing with time-varying exposures with possible time-varying *G-E* association, we need to include the more flexible model as a component for shrinkage estimation in order to control bias.

6. Discussion

The current study, as an extension of the cross-sectional outcome work by Sánchez *et al*. [26], has multiple strengths. It further shows the advantages of using LV models to reduce dimensionality of correlated exposure variables and longitudinal outcome. It limits the number of tests to be conducted and boosts power by using a single integrated model. It also has a cohesive interpretation if the existence of LV is natural, as is the case for blood and bone lead biomarkers in our example. Furthermore, it takes into account the possibility of mutual dependence between gene, exposures, and outcome. By using shrinkage estimation, we avoid testing each assumption separately and still obtain estimates that retain good operating characteristics. The most important contribution is to posit a framework that allows time-varying interaction and time-varying association. When the *G-E* dependence varies over time, a shrinkage factor that accounts for heterogeneity of the *G-E* association is needed.

Further extensions to accommodate semi-parametric or non-parametric smoothing terms instead of the highly parametric longitudinal model we have considered seem a natural continuation of our work. The normality assumption on the random effects and errors can also be relaxed to incorporate non-normal data. Moreover, the outcome can be multivariate longitudinal [47], and one may add another layer of measurement models for the outcome at each time point.

It is also important to consider how to define genetic subgroups to meaningfully stratify on genetic risk for a given outcome–exposure association. For the methods presented, genetic strata need to be determined a priori, for example, through the quantiles of a combined risk index from a pathway or a gene region. In ELEMENT, a very interesting question is how to model the mother-child pair of genes together [6]. However, the limited sample size will always be an issue if one is trying to have multiple genetic subgroups. Whether an LV model can help reduce dimensionality of a correlated gene space is also something to be explored in future studies.

Acknowledgements

This research was supported by ES20811, P30 ES017885, P01 ES022844-01, and P20 ES018171 from NIEHS; R834800 and 83543601 from EPA; and DMS 1406712 from NSF. The authors thank ELEMENT investigators for proving data for the example, as well as the following NIEHS grants that supported data collection: K23ES000381; P01 ES012874; P42 ES05947; R01 ES013744; R01 ES014930; R01 ES007821.

References

1. Afeiche M, Peterson KE, Sánchez BN, Cantonwine D, Lamadrid-Figueroa H, Schnaas L, Ettinger AS, Hernández-Avila M, Hu H, Téllez-Rojo MM. Prenatal lead exposures and weight of 0 to 5-year-old children in Mexico City. *Environmental Health Perspectives* 2011; **119**(10):1436–1441.
2. Fortenberry GZ, Meeker JD, Sánchez BN, Barr DB, Panuwet P, Bellinger D, Schnaas L, Solano-González M, Ettinger AS, Hernandez-Avila M, Hu H, Tellez-Rojo MM. Urinary 3, 5, 6-trichloro-2-pyridinol (TCPY) in pregnant women from Mexico

City: Distribution, temporal variability, and relationship with child attention and hyperactivity. *International Journal of Hygiene and Environmental Health* 2014; **217**(2):405–412.
3. Zhang A, Hu H, Sánchez BN, Ettinger AS, Park SK, Cantonwine D, Schnaas L, Wright RO, Lamadrid-Figueroa H, Tellez-Rojo MM. Association between prenatal lead exposure and blood pressure in children. *Environmental Health Perspectives* 2012; **120**(3):445–450.
4. Henn BC, Kim J, Wessling-Resnick M, Téllez-Rojo MM, Jayawardene I, Ettinger AS, Hernández-Avila M, Schwartz J, Christiani DC, Hu H, Wright RO. Associations of iron metabolism genes with blood manganese levels: a population-based study with validation data from animal models. *Environmental Health* 2011; **10**(1):1–11.
5. Kordas K, Ettinger AS, Bellinger DC, Schnaas L, Téllez-Rojo MM, Hernández-Avila M, Hu H, Wright RO. A Dopamine Receptor (DRD2) but Not Dopamine Transporter (DAT1) gene polymorphism is associated with neurocognitive development of Mexican preschool children with lead exposure. *The Journal of Pediatrics* 2011; **159**(4):638–643.
6. Cantonwine D, Hu H, Téllez-Rojo MM, Sánchez BN, Lamadrid-Figueroa H, Ettinger AS, Mercado-García A, Hernandez-Avila M, Wright RO. HFE gene variants modify the association between maternal lead burden and infant birthweight: A prospective birth cohort study in Mexico City, Mexico. *Environmental Health* 2010; **9**:43.
7. Pilsner JR, Hu H, Wright RO, Kordas K, Ettinger AS, Sánchez BN, Cantonwine D, Lazarus AL, Cantoral A, Schnaas L, Téllez-Rojo MM, Hernández-Avila M. Maternal MTHFR genotype and haplotype predict deficits in early cognitive development in a lead-exposed birth cohort in Mexico City. *The American Journal of Clinical Nutrition* 2010; **92**(1): 226–234.
8. Hopkins MR, Ettinger AS, Hernandez-Avila M, Schwartz J, Téllez-Rojo MM, Lamadrid-Figueroa H, Bellinger D, Hu H, Wright RO. Variants in iron metabolism genes predict higher blood lead levels in young children. *Environmental Health Perspectives* 2008; **116**(9):1261–1266.
9. Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. *Epidemiologic Reviews* 1998; **20**(2):137–147.
10. Kraft P, Hunter D. Integrating epidemiology and genetic association: the challenge of gene–environment interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2005; **360**(1460):1609–1616.
11. Hunter DJ. Gene–environment interactions in human diseases. *Nature Reviews Genetics* 2005; **6**(4):287–298.
12. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nature Reviews Genetics* 2010; **11**(4): 259–272.
13. Khoury MJ, Wacholder S. Invited commentary: From genome-wide association studies to gene-environment-wide interaction studies–challenges and opportunities. *American Journal of Epidemiology* 2009; **169**(2):227–230.
14. Gardner RM, Kippler M, Tofail F, Bottai M, Hamadani J, Grandér M, Nermell B, Palm B, Rasmussen KM, Vahter M. Environmental exposure to metals and children's growth to age 5 years: a prospective cohort study. *American Journal of Epidemiology* 2013; **177**(12):1356–1367.
15. Bandeen-Roche K, Glass TA, Bolla KI, Todd AC, Schwartz BS. The longitudinal association of cumulative lead dose with cognitive function in community-dwelling older adults. *Epidemiology* 2009; **20**(6):831–839.
16. Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. *American Journal of Epidemiology* 2009; **169**(4):497–504.
17. 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.
18. Ko YA, Saha-Chaudhuri P, Park SK, Vokonas PS, Mukherjee B. Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data. *Genetic Epidemiology* 2013; **37**(6):581–591.
19. Muthén B. Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology* 1997; **27**(1):453–480.
20. Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC: Boca Raton, FL, 2004.
21. Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P, White RF. Statistical methods for the evaluation of health effects of prenatal mercury exposure. *Environmetrics* 2003; **14**(2):105–120.
22. Sánchez BN, Budtz-Jørgensen E, Ryan LM, Hu H. Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association* 2005; **100**(472):1443–1455.
23. Dhungana P, Eskridge KM, Baenziger PS, Campbell BT, Gill KS, Dweikat I. Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines. *Crop Science* 2007; **47**(2): 477–484.
24. Rathouz PJ, Van Hulle CA, Rodgers JL, Waldman ID, Lahey BB. Specification, testing, and interpretation of gene-by-measured-environment interaction models in the presence of gene–environment correlation. *Behavior Genetics* 2008; **38**(3):301–315.
25. Javaras KN, Hudson JI, Laird NM. Fitting ACE structural equation models to case-control family data. *Genetic Epidemiology* 2010; **34**(3):238–245.
26. Sánchez BN, Kang S, Mukherjee B. A latent approach to study gene-environment interactions in the present of multiple correlated exposures. *Biometrics* 2012; **68**(2):466–476.
27. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 2005; **92**(2):399–418.
28. 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.
29. Chen YH, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association* 2009; **104**(485):220–233.
30. Weuve J, Sánchez BN, Calafat AM, Schettler T, Green RA, Hu H, Hauser R. Exposure to phthalates in neonatal intensive care unit infants: urinary concentrations of monoesters and oxidative metabolites. *Environmental Health Perspectives* 2006; **114**(9):1424–1431.

31. Budtz-Jørgensen E, Debes F, Weihe P, Grandjean P. Structural equation models for meta-analysis in environmental risk assessment. *Environmetrics* 2010; **21**(5):510–527.
32. Sánchez BN, Budtz-Jørgensen E, Ryan LM. An estimating equations approach to fitting latent exposures models with longitudinal health outcomes. *The Annals of Applied Statistics* 2009; **3**:830–856.
33. Garcia-Closas M, Rothman N, Figueroa JD, Prokunina-Olsson L, Han SS, Baris D, Jacobs EJ, Malats N, De Vivo I, Albanes D, Purdue MP, Sharma S, Fu YP, Kogevinas M, Wang Z, Tang W, Tardón A, Serra C, Carrato A, García-Closas R, Lloreta J, Johnson A, Schwenn M, Karagas MR, Schned A, Andriole Jr G, Grubb III R, Black A, Gapstur SM, Thun M, Diver WR, Weinstein SJ, Virtamo J, Hunter DJ, Caporaso N, Landi MT, Hutchinson A, Burdett L, Jacobs KB, Yeager M, Fraumeni Jr JF, Chanock SJ, Silverman DT, Chatterjee N. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Research* 2013; **73**(7):2211–2220.
34. Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 1993; **58**(4):525–543.
35. Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 1989; **105**(3):456–466.
36. Muthén B. Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In *Handbook of Quantitative Methodology for the Social Sciences*. Sage: Newbury Park, CA, 2004.
37. Wen X, Kleinman K, Gillman MW, Rifas-Shiman SL, Taveras EM. Childhood body mass index trajectories: modeling, characterizing, pairwise correlations and socio-demographic predictors of trajectory characteristics. *BMC Medical Research Methodology* 2012; **12**:38.
38. Bollen KA. *Structural Equations With Latent Variables*. Wiley: New York, 1989.
39. Rosseel Y. lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 2012; **48**(2):1–36.
40. Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* 2005; **29**(2):108–127.
41. Hanson EH, Imperatore G, Burke W. HFE gene and hereditary hemochromatosis: a HuGE review. *American Journal of Epidemiology* 2001; **154**(3):193–206.
42. Gonzalez-Cossio T, Peterson KE, Sanin LH, Fishbein E, Palazuelos E, Aro A, Hernandez-Avila M, Hu H. Decrease in birth weight in relation to maternal bone-lead burden. *Pediatrics* 1997; **100**(5):856–862.
43. Hu H. Bone lead as a new biologic marker of lead dose: recent findings and implications for public health. *Environmental Health Perspectives* 1998; **106**(Suppl 4):961–967.
44. Park SK, Hu H, Wright RO, Schwartz J, Cheng Y, Sparrow D, Vokonas PS, Weisskopf MG. Iron metabolism genes, low-level lead exposure, and QT interval. *Environmental Health Perspectives* 2009; **117**(1):80–85.
45. Raghunathan TE, Solenberger P, Van Hoewyk J. *IVEware: Imputation and Variance Estimation Software User Guide*. Survey Methodology Program, University of Michigan: Ann Arbor, Michigan, 2002.
46. Little RJA, Rubin DB. *Statistical Analysis with Missing Data, 2nd Edition*. Wiley: New York, 2002.
47. Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* 2000; **56**(4):1047–1054.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.