Supplementary Material for Novel Likelihood Ratio Tests for Screening Gene-Gene and Gene-Environment Interactions with Unbalanced Repeated-Measures Data

YI-AN KO[1,*], PARAMITA SAHA-CHAUDHURI[2], SUNG KYUN PARK[3], PANTEL S. VOKONAS[4,5], BHRAMAR MUKHERJEE[1]

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

[2] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

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

[4] VA Normative Aging Study, Veterans Affairs Boston Healthcare System, Boston, MA 02130, USA

[5] Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA

**email:* yianko@umich.edu

Sensitivity of Using the Empirical Variance Estimate for LRT-CM

We studied the sensitivity of using the empirical variance estimates by comparing the power and type I error under true and misspecified correlation structures across a variety of commonly used correlation structures (e.g., compound symmetric, autoregressive, unstructured). Table S.1 shows the simulation results of the proposed LRT-CM method with and without a misspecified correlation structure under compound symmetric (CS) and autoregressive (1) (AR(1)) correlation structures.

Overall, the power of the tests decreases when the assumed correlation structure is more complicated (i.e., more parameters need to be estimated) than the true underlying correlation structure. In our simulation setting, the power of LRT-CM decreases by 16% (at most) when assuming an AR(1) correlation while the true correlation structure is a CS. In contrast, the power of the LRT-CM is not affected as much (<6% in our simulation setting) when the assumed correlation structure requires fewer parameters to be estimated than the true underlying correlation structure. Concerning the type I error under misspecified correlation structure, we again investigated the type I error under two null hypotheses as Figure 2: (1) no interaction with the presence of main effects and (2) no interaction without the presence of main effects. Under the null hypothesis of (2), the estimates can become quite unstable for models (a)–(d). The type I error can be inflated or deflated under misspecified correlation structure but always remain <10% in our simulations.

Estimation for Tukey's Row-Column Model in Two-Step Regression

We can express the interaction term $\theta R_i C_j + \lambda_i C_j + R_i \eta_j$ in the model as $\gamma_{ij} = (\theta R_i + \lambda_i) C_j + R_i \eta_j$ or $(\theta C_j + \eta_j) R_i + \lambda_i C_j$. If we regress the residuals $\boldsymbol{r}_{ijk}$ after removing the additive main effects (from a saturated model fit) on $\hat{C}_j$ and $\hat{R}_i$ (again without intercept) separately:

$$\boldsymbol{r}_{ijk} = u_i \hat{C}_j \mathbf{1}_{n_{ijk}} + \boldsymbol{\epsilon}_{ijk}, \tag{1}$$

$$\boldsymbol{r}_{ijk} = v_j \hat{R}_i \mathbf{1}_{n_{ijk}} + \boldsymbol{\epsilon}_{ijk}, \tag{2}$$

we have $\hat{u}_i = \tilde{\theta} \hat{R}_i + \tilde{\lambda}_i$, and $\hat{v}_j = \tilde{\theta} \hat{C}_j + \tilde{\eta}_j$. Model (d) has a total of $I + J + 1$ interaction parameters. Together with four sum-to-zero identifiability constraints, $I + J - 3$ parameters (i.e., $\lambda_1 \ldots, \lambda_{I-2}, \eta_1, \ldots, \eta_{J-2}$) are left to be estimated. By (1) and (2), we have $(I - 1) + (J - 1)$ equations, which are sufficient for estimating the $I + J - 3$ parameters. After obtaining $\hat{u}_i$ and $\hat{v}_j$ from (1) and (2), each $\hat{\lambda}_i$ and $\hat{\eta}_j$ can be calculated using the constraints. Finally, we estimate θ by regressing the residuals from the second step, $\boldsymbol{s}_{ijk} = \boldsymbol{r}_{ijk} - \hat{R}_i \hat{\eta}_j \mathbf{1}_{n_{ijk}} - \hat{\lambda}_i \hat{C}_j \mathbf{1}_{n_{ijk}}$, on $\hat{R}_i \hat{C}_j$,

$$\boldsymbol{s}_{ijk} = \theta \hat{R}_i \hat{C}_j \mathbf{1}_{n_{ijk}} + \boldsymbol{\varepsilon}_{ijk},$$

where $\boldsymbol{\varepsilon}_{ijk} \sim N(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}})$. Again, $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}$ can be a user-defined covariance structure based on model fitting criterion.

Comparison with Other Exisiting GGI/GEI Methods

The existing GGI or GEI methods for handling (longitudinal) continuous traits are very limited. Barhdadi and Dubé [1] have applied Tukey's and Mandel's models as well as AMMI models to testing GGI effects on quantitative traits for unbalanced data. They reduced data to cell means and applied F tests that assume equal variance of all cell means as described in the original papers of Tukey [2] and Mandel [3]. The likelihood ratio test proposed by Johnson and Graybill [4] was used for GGI tests with AMMI models, which is also based on single observation per cell. Despite these complex classical models, a saturated model for interaction is commonly used for testing GGI and GEI in practice for its computational simplicity and flexibility.

We generated interaction data in the same simulation setting as described in the main text (unbalanced correlated data in 3×3 table settings) and applied the GGI tests summarized in Barhdadi and Dubé [1] for Tukey's, Mandel's, and AMMI models (any within-subject correlation is ignored). Figure S.2 shows type I error (left panel) and power (right panel) for each of the five multiplicative models using tests in Barhdadi and Dubé and our proposed tests (LRT-CM and LRT-PB) under the same simulation settings as described in the section of Simulation Settings in the main text. As expected, the tests in Barhdadi and Dubé [1] assuming balanced data structure and not accounting for within-subject correlations yield inflated type I error (especially for Tukey's and Mandel's models) and low power, compared to our proposed methods. For example, when the simulation model is AMMI1 with $\sigma_b^2 = \sigma_e^2 = 8$, AMMI1 has 65% and 69% power for detecting interactions using our proposed LRT-CM and LRT-PB, respectively; whereas AMMI1 using the test by Barhdadi and Dubé only has 8% power (far right column).

Stratified Analysis of GEI in the NAS Data by Baseline Age

To further investigate the potential three-way interaction (age contributions to *HFE* $\times$ Lead interaction), we performed stratified analysis for by baseline age: one for those who started the study at age $<$ 66 years old (N=316) and the other one with those who started at age $\geq$ 66 years old (N=355). We then analyzed the two subsets separately. The results (p-values) using models (a)–(e) are shown in Table S.4. The results indicate that *HFE* $\times$ Lead interaction was found for the older group of participants but not for the younger group. The stratified analysis results may indicate some evidence of three-way (age-dependent) interaction.

References

[1] Barhdadi, A. and Dubé, M. (2010). Testing for gene-gene interaction with AMMI models. *Statistical Applications in Genetics and Molecular Biology* **9**, 1-27.

[2] Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics* **5**, 232-242.

[3] Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association* **56**, 878-888.

[4] Johnson, D. and Graybill, F. (1972). An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association* **67**, 862-868.

TABLE S.1. Power and type I error of the LRT-CM method with and without a misspecified correlation structure. Two covariance structures were compound symmetric and autoregressive(1) correlation ($\sigma^2 = 16, \rho = 0.5$).

	Corr. Structure		Model				
	True	Assumed	(a) T1	(b) MC	(c) MR	(d) TRC	(e) AMMI1
Power	CS	CS	0.592	0.947	0.747	0.825	0.651
	CS	AR(1)	0.528	0.924	0.660	0.750	0.544
	AR(1)	AR(1)	0.693	0.973	0.858	0.921	0.779
	AR(1)	CS	0.722	0.977	0.888	0.940	0.830
Type I Error (Additive)	CS	CS	0.054	0.052	0.054	0.053	0.055
	CS	AR(1)	0.044	0.034	0.035	0.024	0.027
	AR(1)	AR(1)	0.055	0.057	0.053	0.058	0.056
	AR(1)	CS	0.074	0.079	0.082	0.087	0.088
Type I Error (Null)	CS	CS	0.127	0.087	0.081	0.023	0.057
	CS	AR(1)	0.091	0.057	0.050	0.012	0.027
	AR(1)	AR(1)	0.125	0.083	0.080	0.023	0.056
	AR(1)	CS	0.147	0.116	0.098	0.037	0.088

CS = compound symmetric; AR(1) = autoregressive(1)

TABLE S.2. Percent bias and mean squared error (MSE) corresponding to the interaction parameter estimates from Tukey's 1-df model (θ) and AMMI1 model (d_1) using a two-step regression procedure under compound symmetric and autoregressive(1) correlation structures (both with $\rho = 0.5$)

			Assumed Correlation Structure for Analysis				
Parameter	True	σ^2	CS	AR(1)	ARH	UN	IND
Percent Bias (%)							
θ	CS	1	-0.2	-0.2	-0.2	-0.2	-0.2
		4	0.2	0.3	0.3	0.2	0.3
		8	0.8	0.8	0.8	0.8	0.7
	AR(1)	1	0.1	0.1	0.1	0.1	0.1
		4	0.7	0.7	0.7	0.7	0.7
		8	-0.4	-0.4	-0.4	-0.4	-0.4
d_1	CS	1	1.2	1.2	1.2	1.2	1.2
		4	3.0	3.0	3.0	3.0	3.1
		8	6.8	6.8	6.9	6.8	7.0
	AR(1)	1	1.0	1.0	1.0	1.0	1.0
		4	1.9	1.9	1.9	1.9	1.8
		8	3.6	3.3	3.3	3.3	3.7
MSE							
θ	CS	1	0.090	0.090	0.090	0.090	0.090
		4	0.182	0.185	0.185	0.183	0.185
		8	0.263	0.268	0.268	0.263	0.270
	AR(1)	1	0.078	0.077	0.077	0.077	0.078
		4	0.163	0.162	0.162	0.162	0.163
		8	0.239	0.239	0.239	0.239	0.239
d_1	CS	1	0.101	0.102	0.103	0.101	0.104
		4	0.199	0.204	0.204	0.200	0.206
		8	0.278	0.282	0.282	0.278	0.284
	AR(1)	1	0.096	0.093	0.093	0.093	0.095
		4	0.178	0.175	0.176	0.176	0.179
		8	0.252	0.250	0.250	0.251	0.253

CS = compound symmetric; AR(1) = autoregressive(1); ARH = autoregressive heterogeneous; UN = unstructured; IND = independence
True $\theta = d_1 = 1$

TABLE S.3. Estimated interaction matrices from fitting a saturated model (adjusted for baseline age, time, and squared time) and the corresponding singular value decompositions: $\hat{\mathbf{\Gamma}}_{G\times E}$ for gene-environment (*HFE* $\times$ Lead) interaction analysis based on the Normative Aging Study data

$\hat{\mathbf{\Gamma}}_{G\times E}$

1.92	0.77	-2.68
-0.19	1.14	-0.94
-1.73	-1.90	3.63

$\hat{\boldsymbol{A}}_{\text{HFE}}$

-0.59	0.57
-0.20	-0.79
0.79	0.22

$\hat{\boldsymbol{D}}$

5.65	0
0	1.24

$\hat{\boldsymbol{B}}'_{\text{Lead}}$

-0.43	-0.38	0.82
0.69	-0.72	0.03

TABLE S.4. P-values corresponding to different tests for GEI between *HFE* genotypes and tibia lead levels in the Normative Aging Study stratified by baseline age at the time of recruitment are reported. LRT-CM and LRT-PB stand for the two likelihood ratio tests based on cell means and al mixed-effects regression model, respectively. The model adjusts for baseline age (years), time since baseline, and squared time. For LRT-CM, the residuals from the adjusted model were used to form cell means corresponding to $G \times E$ cross-tables.

		Baseline Age <66		Baseline Age ≥ 66	
Model	Hypothesis	LRT-CM	LRT-PB	LRT-CM	LRT-PB
Model (a)	$H_0 : \theta = 0$	0.054	0.208	0.001	0.001
Model (b)	$H_0 : \lambda_i = 0$ (Lead)	0.143	0.080	0.003	0.001
Model (c)	$H_0 : \eta_j = 0$ (*HFE*)	0.133	0.142	0.001	0.002
Model (d)	$H_0 : \theta = \lambda_i = \eta_j = 0$	0.234	0.184	0.002	$<$.0001
Model (e)	$H_0 : d_1 = 0$	$<$0.10	0.250	$<$0.005	0.001
Saturated Model	*HFE* $\times$ Lead	NA	0.284	NA	0.002

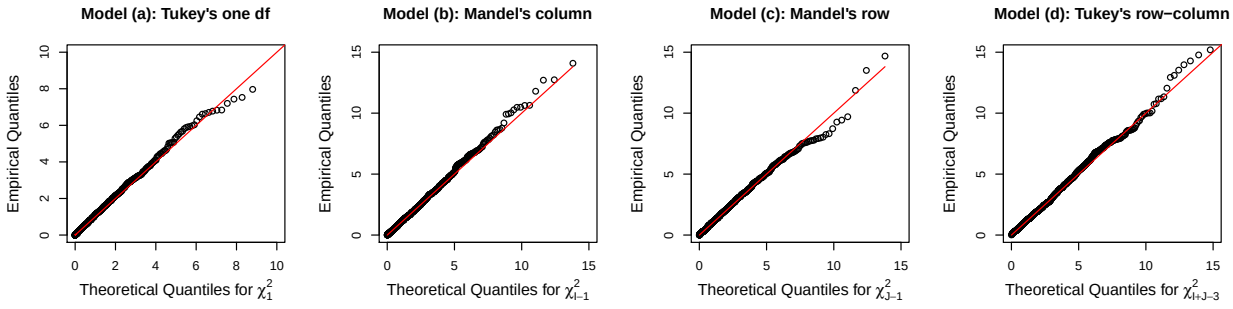
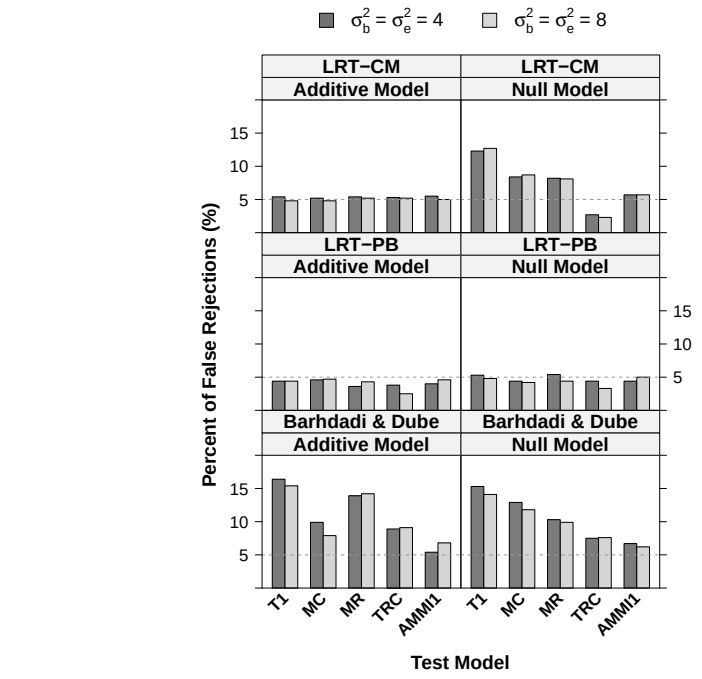
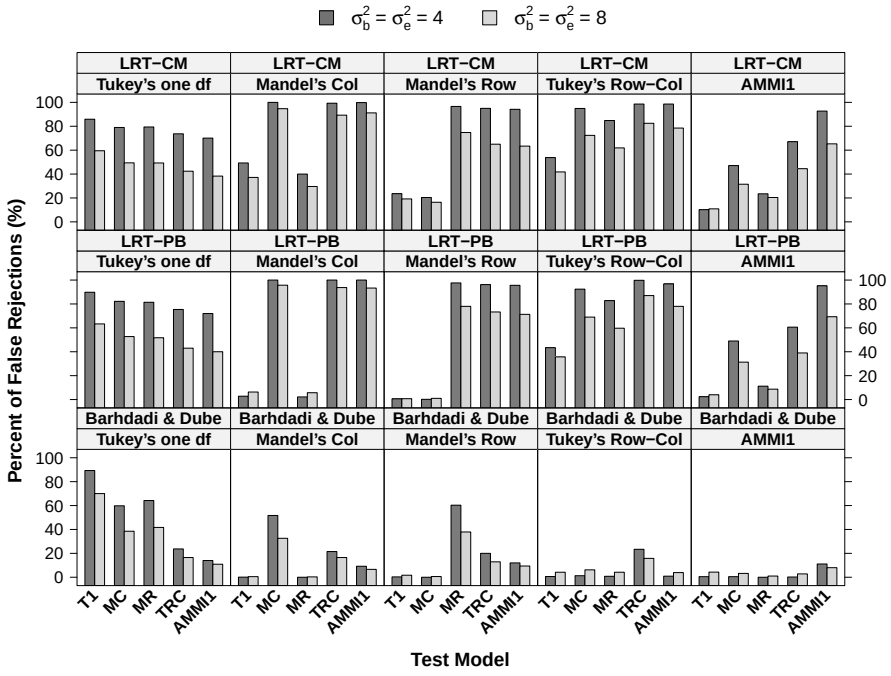


Fig. S.1. Comparison of empirical quantiles of the likelihood ratio test (LRT) statistics to the corresponding theoretical quantiles of chi-squares under the null hypothesis based on $I \times J$ cell means. The LRT statistic follows a chi-square distribution with df $= 1, I-1, J-1$, and $I+J-3$ for models (a), (b), (c), and (d), respectively ($I = J = 3$).



(a)



(b)

Fig. S.2. (a) Type I error and (b) percentage of interactions detected by each of the five multiplicative models using tests in Barhdadi and Dubé (2010) and the proposed methods in the same simulation settings as described in the section of Simulation Settings. The top label within each box represents the true simulation model. The horizontal-axis labels indicate the tests carried out.

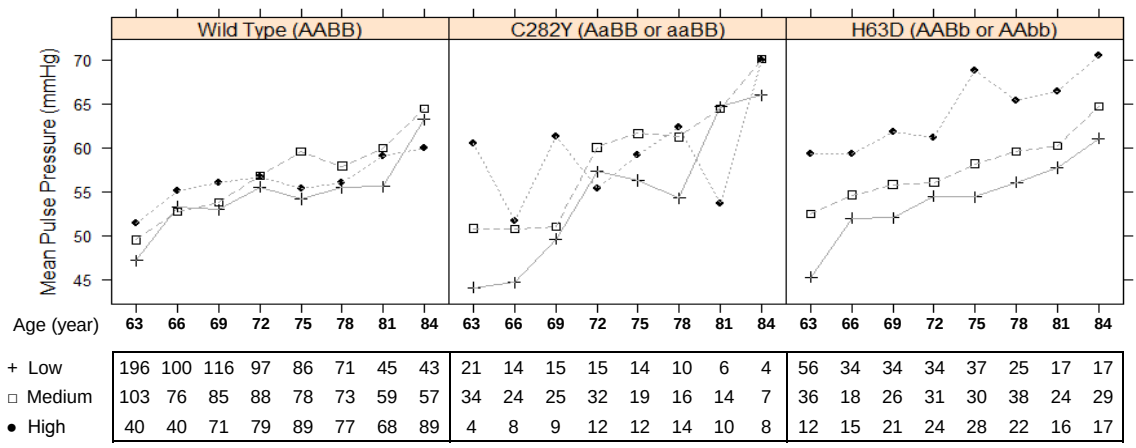


Fig. S.3. Cell mean of pulse pressure and number of observations (shown in table below the graph) for three genotypes of the *HFE* gene and lead exposure levels (Low, Medium, High) across eight age intervals in the Normative Aging Study

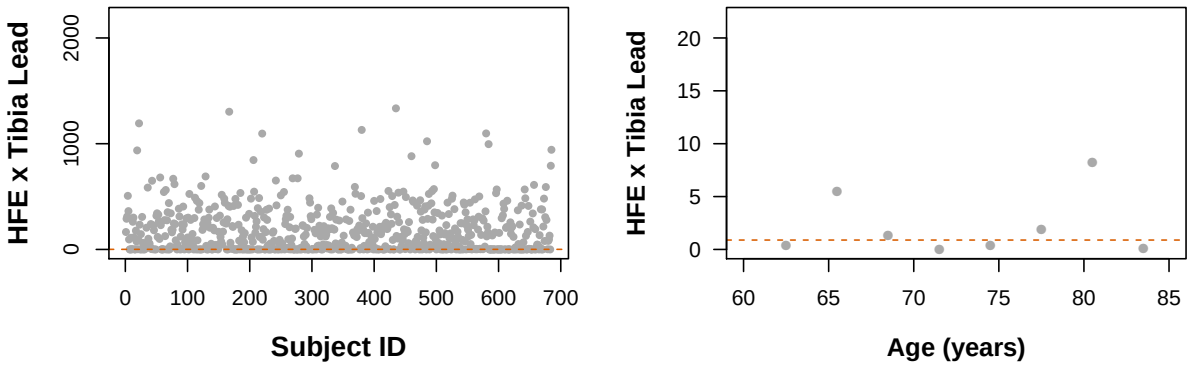


Fig. S.4. Subject-specific contributions and age-specific contributions to the *second* interaction factor in the *HFE* $\times$ Lead interaction based on the Normative Aging Study data