

Web-Based Supporting Materials for Exposure Enriched Outcome Dependent Designs for Longitudinal Studies of Gene-Environment Interaction

BY ZHICHAO SUN, BHRAMAR MUKHERJEE, JASON P. ESTES, PANTEL S. VOKONAS, SUNG KYUN PARK

1 Expanded Methods

1.1 Unweighted Uncorrected Likelihood

Regardless of the sampling mechanism, one can naively perform the analysis for a standard prospective cohort, and make inference on parameters of interest in the linear mixed model by equating the derivative of the unweighted uncorrected log-likelihood (UUL) to zero:

$$\sum_{i:S_i=1} \frac{\partial \log f(Y_i|X_i; \beta, \sigma)}{\partial \beta} = 0$$

Solutions to this equation yield the maximum likelihood estimates of β (or σ). One problem of this naive analysis is that there is no guarantee for consistent estimates when sample selection in Phase II is in relation to the outcome.

1.2 Inverse Probability Weighted Likelihood

To draw valid inference in a two-phase design, we consider the inverse probability weighted likelihood (IPWL), a modification of complete-case analysis that differentially weights subjects to adjust for the selection bias.

In particular, subject-specific selection probability $P(S_i = 1|Y_i, X_i)$ for the entire cohort $i = 1, \dots, N$ can be obtained by matching the stratum-specific selection probability with identified personal stratum membership. Based upon information on complete-cases, consistent estimators of β (or σ) can be derived by solving the estimation equation

$$\sum_{i:S_i=1} \frac{\partial \log f(Y_i|X_i; \beta, \sigma)}{\partial \beta} \cdot [P(S_i = 1|Y_i, X_i)]^{-1} = 0$$

Here, the contribution to the score function from a single subject in Phase II is weighted by the inverse of its sampling probability (Robins et al., 1994). When a constant sampling probability is assigned, such as n/N in random sampling, it is easy to show that the IPWL becomes equivalent to the UUL.

1.3 Complete-Case Conditional Likelihood

To adjust for biased sampling from outcome dependent sampling, Schildcrout et al. (2013) developed an ascertainment corrected maximum likelihood for inference. In their analysis, subjects in the set of complete-cases

contribute to the likelihood by a conditional probability of the outcome vector given being sampled in Phase II, $P(Y_i|X_i, S_i = 1; \beta, \sigma)$. By Bayes' theorem, its likelihood function, which we refer to as the complete-case conditional likelihood (CCL), can be derived as:

$$L^C(\beta, \sigma) = \prod_{i:S_i=1} f(Y_i|X_i, S_i = 1; \beta, \sigma) = \prod_{i:S_i=1} \frac{P(S_i = 1|Y_i, X_i)f(Y_i|X_i; \beta, \sigma)f(X_i)}{P(S_i = 1|X_i; \beta, \sigma)f(X_i)} = \prod_{i:S_i=1} \frac{\pi(q_i) \cdot f(Y_i|X_i; \beta, \sigma)}{P(S_i = 1|X_i; \beta, \sigma)}$$

where subject-specific contribution to the CCL is composed of three terms: $f(Y_i|X_i; \beta, \sigma)$ the multivariate density in the response model, $P(S_i = 1|X_i; \beta, \sigma)$ the subject-specific correction that adjusts for the biased sampling, and $P(S_i = 1|Y_i, X_i) = \pi(q_i)$ the subject-specific selection probability determined by observed q_i , which is functionally independent of parameters β and σ .

Under the normality assumption of $Y_i|X_i$ in the response model, the sampling variable $Q_i|X_i$ should also follow a normal distribution with its mean and covariance indexed by parameters β and σ , i.e., $Q_i|X_i \sim \mathcal{N}(\mu_{q_i}(\beta), \Sigma_{q_i}(\sigma))$. Therefore, the correction term for subject i can be computed as a weighted average of stratum-specific selection probabilities across all strata, $P(S_i = 1|X_i; \beta, \sigma) = \sum_{k=1}^K \pi(R^k)P(Q_i \in R^k|X_i; \beta, \sigma)$, which ensures the expression of CCL in a closed form. Score functions of the CCL with respect to parameters β and σ can be calculated analytically as detailed in Schildcrout et al. (2013). One can obtain the CCL estimates by solving the corresponding score equations using the Newton-Raphson algorithm, and the covariance matrix by the inverse of the numerical derivative of the score function.

2 Supplemental Table

Table 1: Average standard errors and empirical standard errors of GxE interaction and joint exposure effects, using fixed values from population distribution of \hat{a}_{0i} as cutpoints in construction of the sampling strata. Results are based on 1000 replicates, each including a cohort of 1000 subjects from which 250 are selected for retrospective genotyping. Unbalanced longitudinal outcome is considered, with a monotone missing pattern of 10% random dropouts at each follow-up visit and up to 5 measurements for each subject. Genotype assumed to be independent of personal exposure, with a prevalence of 0.2. $\beta_E = -1.5$ and $\beta_{GE} = -1.5$.

Sampling scheme	Measure	UUL		IPWL		CCL		FCL	
		β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$
Random	Average SE	1.27	1.21	-	-	-	-	-	-
	Empirical SE	1.33	1.26	-	-	-	-	-	-
E-enriched	Average SE	0.98	0.93	-	-	-	-	0.92	0.85
	Empirical SE	0.98	0.93	-	-	-	-	0.92	0.86
OLS-based	Average SE	1.18	1.01	1.60	1.38	0.97	0.92	0.94	0.91
	Empirical SE	1.14	1.00	1.61	1.55	0.93	0.89	0.89	0.87
E + OLS	Average SE	1.36	1.28	1.32	1.15	0.81	0.74	0.74	0.72
	Empirical SE	1.32	1.19	1.41	1.34	0.76	0.70	0.70	0.68
BLUP-based	Average SE	1.09	0.98	1.53	1.29	0.85	0.82	0.82	0.76
	Empirical SE	1.05	0.95	1.57	1.48	0.81	0.77	0.77	0.71
E + BLUP	Average SE	1.24	1.11	1.35	1.17	0.71	0.68	0.69	0.64
	Empirical SE	1.17	1.07	1.43	1.36	0.67	0.63	0.68	0.63

Table 2: Average standard errors and empirical standard errors of GxE interaction and joint exposure effects, using quantiles of the empirical distribution of \hat{a}_{0i} estimated in the full cohort as cutpoints in construction of the sampling strata. Results are based on 1000 replicates, each including a cohort of 1000 subjects from which 250 are selected for retrospective genotyping. Unbalanced longitudinal outcome is considered, with a monotone missing pattern of 10% random dropouts at each follow-up visit and up to 5 measurements for each subject. Genotype assumed to be independent of personal exposure, with a prevalence of 0.2. $\beta_E = -1.5$ and $\beta_{GE} = -1.5$.

Sampling scheme	Measure	UUL		IPWL		CCL		FCL	
		β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$	β_{GE}	$\beta_E + \beta_{GE}$
Random	Average SE	1.27	1.21	-	-	-	-	-	-
	Empirical SE	1.33	1.26	-	-	-	-	-	-
E-enriched	Average SE	0.98	0.93	-	-	-	-	0.92	0.85
	Empirical SE	0.98	0.93	-	-	-	-	0.92	0.86
OLS-based	Average SE	1.17	1.01	1.61	1.39	0.96	0.92	0.93	0.90
	Empirical SE	1.13	0.98	1.70	1.61	0.93	0.89	0.90	0.85
E + OLS	Average SE	1.32	1.24	1.33	1.16	0.77	0.73	0.74	0.71
	Empirical SE	1.28	1.21	1.45	1.39	0.76	0.70	0.74	0.68
BLUP-based	Average SE	1.08	0.97	1.51	1.27	0.85	0.82	0.82	0.76
	Empirical SE	1.05	0.94	1.59	1.48	0.83	0.78	0.79	0.73
E + BLUP	Average SE	1.23	1.10	1.35	1.17	0.70	0.67	0.68	0.63
	Empirical SE	1.18	1.08	1.43	1.37	0.69	0.65	0.68	0.63

Table 3: Estimated bias for GxExT interaction and joint exposure effect under the three-way GxExT interaction model using different sampling designs and likelihood approaches. Results are based on 1000 replicates, each including a cohort of 5000 subjects from which 500 are selected for retrospective genotyping. Unbalanced longitudinal outcome is considered, with a monotone missing pattern of 10% random dropouts at each follow-up visit and up to 5 measurements for each subject. Genotype assumed to be independent of personal exposure, with a prevalence of 0.2. $\beta_{GET} = -1.5$ and $\beta_E + \beta_{ET} + \beta_{GE} + \beta_{GET} = -4.5$. Estimates biased by at least 10% in bold.

Sampling scheme	Sampling variable	UUL		IPWL		CCL		FCL	
		GxExT	Joint E	GxExT	Joint E	GxExT	Joint E	GxExT	Joint E
Random	-	0.01	0.00	-	-	-	-	-	-
E-enriched	E_i	-0.02	-0.03	-	-	-	-	-0.01	0.00
OLS-based	$\hat{\eta}_{1i}$	3.27	0.99	-0.04	-0.07	0.04	0.01	0.02	0.01
E + OLS	$(E_i, \hat{\eta}_{1i})$	3.42	1.28	0.03	0.01	0.02	0.02	0.00	0.00
BLUP-based	\hat{a}_{1i}	0.33	0.68	-0.12	-0.17	0.02	0.02	0.01	0.01
E + BLUP	(E_i, \hat{a}_{1i})	1.42	1.16	-0.02	0.11	0.02	-0.02	0.00	0.00

3 Supplemental Figures

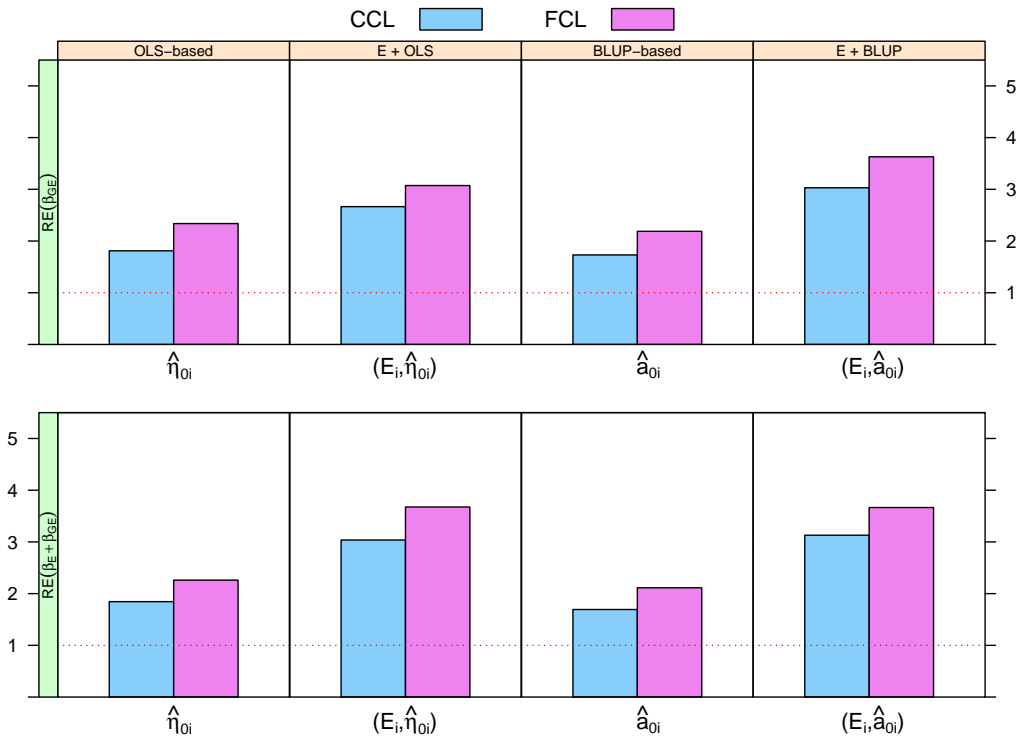


Figure 1: Relative efficiency of GxE interaction and joint exposure effect under different sampling designs in the analysis of CCL and FCL. Results are based on 1000 replicates, each including a cohort of 1000 subjects from which 250 are selected for retrospective genotyping. Balanced longitudinal outcome with 5 measurements for each subject is considered. Genotype assumed to be independent of personal exposure, with a prevalence of 0.2.

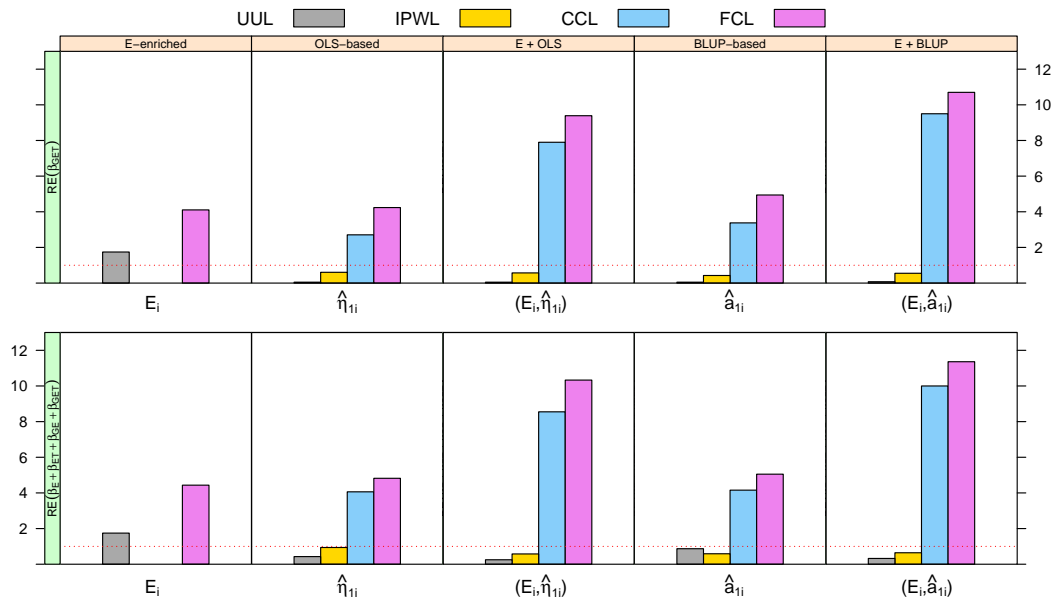


Figure 2: Relative efficiency of GxExT interaction and joint exposure effect under different sampling designs and likelihood approaches. Results are based on 1000 replicates, each including a cohort of 5000 subjects from which 500 are selected for retrospective genotyping. Unbalanced longitudinal outcome is considered, with a monotone missing pattern of 10% random dropouts at each follow-up visit and up to 5 measurements for each subject. Genotype assumed to be independent of personal exposure, with a prevalence of 0.2. $\beta_{GET} = -1.5$ and $\beta_E + \beta_{ET} + \beta_{GE} + \beta_{GET} = -4.5$.