Reducing Selection Bias in Analyzing Longitudinal Health Data with High Mortality Rates

Xian Liu, PhD; Charles C. Engel, Jr., MD, MPH
Han Kang, DR.PH; Kristie L. Gore, PhD

Disclaimer

"The views expressed in this presentation are those of the authors and do not necessarily represent the official policy or position of the Deployment Health Clinical Center, Walter Reed Army Medical Center, Department of Defense or the U.S. Government."

Research Objectives

1. Develop two new generalized linear mixed models to handle dropouts due to high mortality.

2. Demonstrate problems when applying conventional approaches to analyze longitudinal health data.

3. Provide an empirical example to introduce how to apply the new approaches.

Research Significance

1. The research helps scientists understand potential limitations in conventional longitudinal models.

2. The study develops new methods to analyze longitudinal health data with high mortality.

3. The new models provide more accurate health data for policy-makers and scientists.

Hypotheses

1. Non-independence of random errors causes inconsistencies of parameter estimates.

2. The covariance between disturbances of the full and the truncated mixed models is statistically significant.

3. The nonparametric two-stage model is the most appropriate statistical approach to analyze the longitudinal health data.

Model Specifications

1. The conventional 1-step mixed model:

$$\left(Y \middle| S=1\right) = X_1\beta_1 + Z_1\gamma_1 + \varepsilon_1.$$

2. The parametric two-stage model:

$$\Pr\left(S=1\right) = \Phi\left(X_2\beta_2 + Z_2\gamma_2\right)$$

$$Y \middle| S = 1 = X_1\beta_3 + Z_1\gamma_3 + \sigma_{12}\lambda + \varepsilon_3.$$

3. The nonparametric two-step mixed model:

$$\Pr\left(Y > 0 \middle| S=1\right) = \Phi\left(X_1\beta_4 + Z_2\gamma_4\right)$$

$$\log[Y\left(Y > 0 \middle| S=1\right)] = \left(X_1\beta_5 + Z_1\gamma_5 + \varepsilon_5\right).$$

The expected number at time t

$$E\left(\hat{Y} \middle| S=1\right) = \Phi\left(X_1\hat{\beta}_4 + Z_2\hat{\gamma}_4\right)\exp\left(X_1\hat{\beta}_5 + Z_1\hat{\gamma}_5\right)\hat{\xi}.$$

Empirical Example

1. Data Source – The Survey of Asset and Health Dynamics among the Oldest Old (AHEAD), Wave I through Wave VI

2. Measure of functional status – The number of functional limitations (ADL, IADL, etc.)

3. Covariates – Time, veteran status, age, gender, education, ethnicity, and the like.

4. Autoregressive error structure – Repeated / Type=SP in executing SAS PROC.MIXED.

Table 1. Results of three mixed models on number of functional limitations in older Americans: AHEAD Longitudinal Survey (n=8,443)

Explanatory Variables and Other Statistics	Conventional Mixed Model	Parametric 2-step Model[a]	Nonparametric 2-step Model[b]
Fixed Effects:			
Intercpt	5.5045**	5.3967**	1.4515**
Time 0 (1993)	-3.0158**	-2.9079**	-0.4582**
Time 1 (1995)	-0.2583**	-0.1320	0.0028
Time 2 (1998)	0.8780**	0.9613**	0.2348**
Time 3 (2000)	0.9984**	1.0416**	0.2287**
Time 4 (2002)	1.2367**	1.2575**	0.2569**
Veteran status	0.1613	0.1023	0.0292
Age	0.1742**	0.1320**	0.0274**
Female	0.7360**	0.8773**	0.0849**
Education	-0.1665	-0.1519**	-0.0269**
Lambda (λ)		-4.8184**	
Random Effects:			
Spatial power (POW)	0.5651**	0.5295**	0.4571**
Residual	12.3156**	11.5321**	0.4939**
Model Chi-square	13367.1**	16715.9**	6100.3**

* 0.01 < P < 0.05; ** P < 0.01

a These are the results of the second-step mixed model.

b These are the results of the second-step mixed model for those with at least one functional limitation, with the dependent variable being the natural logarithm of the number of functional limitations.

Table 2. Predicted number of functional limitations in older Americans Derived from three random-effects models (n=8,443)

Time Points and Parameters	Observed and Predicted Number of Functional Limitations			
	Observed	Conventional	Parametric	Nonparametric
1993	2.4887	2.4996	2.4759	2.6918
1995	5.1514	5.2571	5.2518	5.1184
1998	6.1378	6.3934	6.3451	6.1197
2000	6.1602	6.5138	6.4254	6.1598
2002	6.3348	6.7521	6.6413	6.3056
2004	4.9608	5.5154	5.3838	4.9088

Note: All predicted values derived from the three mixed models are statistically significant relative to value zero.

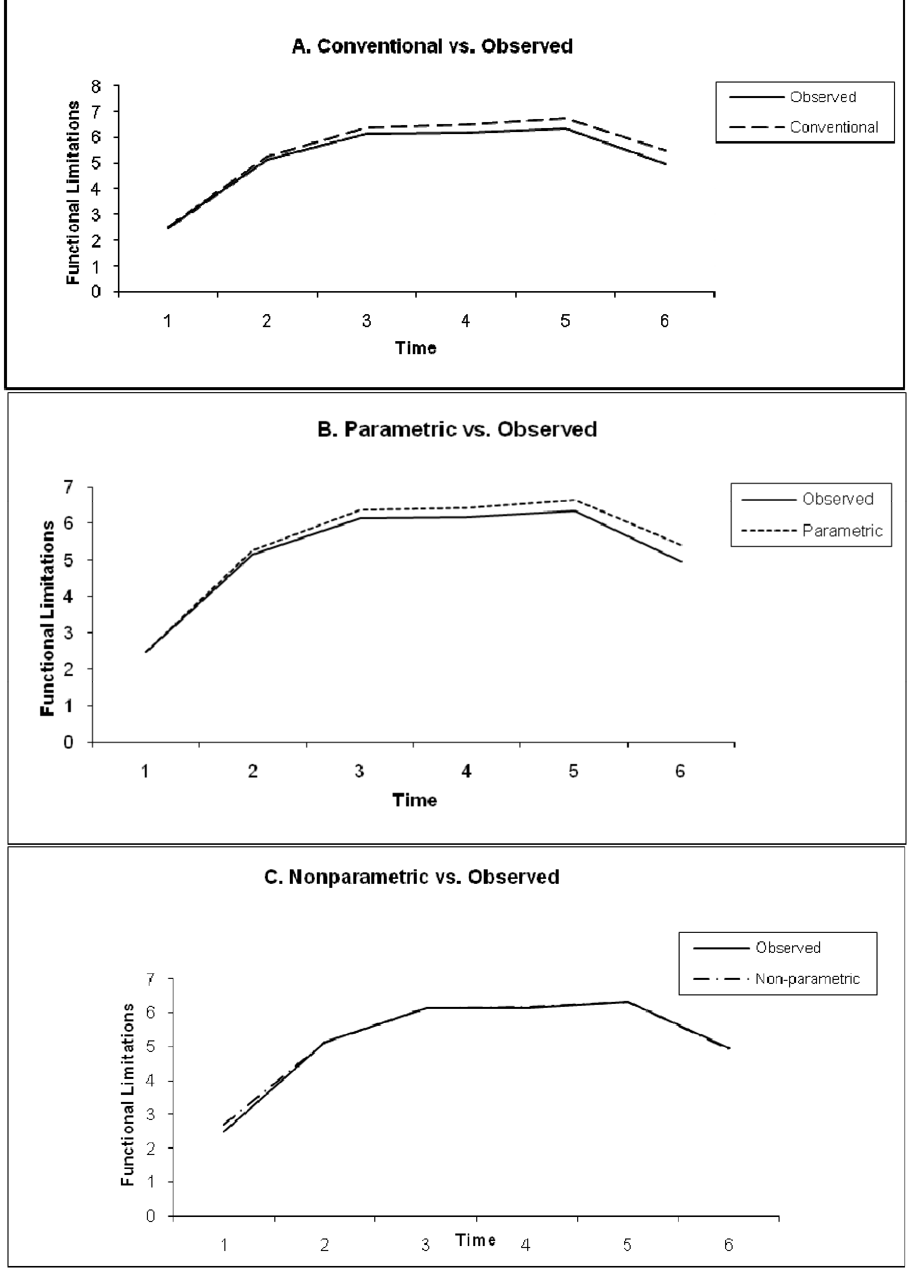

Figure 1. Transitions in number of functional limitations in older Americans:
Growth curves derived from three approaches

Conclusions

1. Direct application of one-step linear random mixed models on longitudinal health data can lead to serious prediction bias.

2. Application of more refined two-stage random-effects models substantially reduces such biases, especially using the nonparametric approach.