Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 9

11-1-2010

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

Xian Liu
Uniformed Services University of the Health Sciences, Bethesda MD and Walter Reed National Military Medical Center, Bethesda MD, Xian.Liu@usuhs.edu

Charles C. Engel
Uniformed Services University of the Health Sciences, Bethesda MD and Walter Reed National Military Medical Center, Bethesda MD, cengel@usuhs.mil

Han Kang
Department of Veterans Affairs, han.kang@mail.va.gov

Kristie L. Gore
Walter Reed National Military, Kristie.gore@med.navy.mil

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 403-413

1538 – 9472/10/$95.00

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

Xian Liu Charles C. Engel
Uniformed Services University of the Health Sciences, Bethesda MD and Walter Reed National Military Medical Center, Bethesda MD

Han Kang
Department of Veterans Affairs, Washington, DC

Kristie L. Gore
Walter Reed National Military Medical Center, Bethesda MD and Uniformed Services University of the Health Sciences, Bethesda MD

Two longitudinal regression models, one parametric and one nonparametric, are developed to reduce selection bias when analyzing longitudinal health data with high mortality rates. The parametric mixed model is a two-step linear regression approach, whereas the nonparametric mixed-effects regression model uses a retransformation method to handle random errors across time.

Key words: Longitudinal data, mortality rates, nonrandom dropouts, selection bias.

Introduction

Analyzing large-scale longitudinal health data poses special challenges to statisticians, demographers and other quantitative methodologists. Most longitudinal surveys collect random and unbiased samples at baseline. Among older persons, however, a considerable proportion of the baseline respondents will not survive to the ensuing phases of investigation. As a result, longitudinal health outcomes are based on several follow-up samples selected by values of the dependent health variable because physically frailer, functionally disabled and environmentally disadvantaged persons are more likely to die. Thus, follow-up data of a longitudinal health survey on these populations often bear little resemblance to the initial sample, making dropouts non-ignorable. Consequently, currently existing longitudinal regression models, such as the random-effects linear regression model, can be highly sensitive to untestable assumptions and inestimable parameters (Hedeker & Gibbons, 2006; Hogan, Roy, & Korkontzelou 2004; Little & Rubin, 2003; Schafer & Graham, 2002).

There is abundant literature devoted to modeling non-ignorable longitudinal missing data in biostatistics (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little, 1995; Little & Rubin, 2003; Robins, Rotnitzky & Zhao, 1995; Yao, Wei & Hogan, 1998). The primary focus of this literature, however, is dropout in clinical trials. Here the missingness is primarily due to reasons other than death and is closely related to outcomes being measured (Schafer & Graham, 2002). In large-scale longitudinal health data for older persons, high death rates are usually the primary reason for dropouts in follow-up waves; in a strict sense, this cannot be simply viewed as missing because the deceased no longer

Xian Liu is Associate Research Professor in the Department of Psychiatry at the F. Edward Hebert School of Medicine, Uniformed Services University of the Health Sciences. Email him at: Xian.Liu@usuhs.edu. Charles C. Engel, Jr. is Associate Chair of the Department of Psychiatry and DoD Director of the Deployment Health Clinical Center at Walter Reed in Bethesda MD. Email him at: cengel@usuhs.mil. Han Kang is Director of Environmental Epidemiology Service, Veterans Health Administration of Department of Veterans Affairs. E-mail him at: han.kang@mail.va.gov. Kristie L. Gore is Director of Research and Program Evaluation at the DoD Deployment Health Clinical Center and Assistant Research Professor in the Department of Psychiatry at the Uniformed Services University of the Health Sciences. E-mail her at: Kristie.gore@med.navy.mil.

possesses any values or characteristics to estimate (Hogan, Roy & Korkontzelou, 2004; Pauler, McCoy &Moinpour, 2003). On the other hand, although assumptions on measurability of the deceased's health outcomes are imperceptible and inappropriate, the influence of high mortality on the distribution of survivors' health data cannot be ignored. When creating a longitudinal model with high death rates, researchers should establish the statistical structure needed to account for the potential lack of independence that often exists among those who have been selected from the survival of the fittest process.

Some researchers have proposed the use of joint modeling, originally developed by Heckman (1979), for longitudinal and survival data that link the health outcomes by means of a common selection factor (Egleston, Scharfstein, Freeman & West, 2006; Fu, Winship & Mare, 2004; Kurland & Heagerty, 2005; Leigh, Ward & Fries, 1993; Pauler, McCoy & Moinpour, 2003; Ratcliffe, Guo & Ten Have, 2004). Given specification of the selection factor, the two responses, survival and longitudinal health outcomes, are thought to be conditionally independent, hence more efficient and less-biased parameter estimates can be obtained from this type of statistical modeling. However, the two-step parametric joint modeling has been criticized because of its considerable dependence on distributional assumptions for the non-ignorable missing data that are impossible to verify (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little & Rubin, 2003; Winship & Mare, 1992). Due to the unique characteristics involved in health transitions among older persons, the restrictive assumptions of this method on the parametric disturbance function can be readily violated, thereby degrading the quality of parameter estimates and model-based prediction.

This research develops two longitudinal regression models to account for the selection bias from high mortality rates, one parametric and one nonparametric. The parametric model is a two-step statistical technique developed as a joint model combining longitudinal and survival data. By contrast, the nonparametric longitudinal model uses a retransformation approach, taking into account the missing data mechanism by assuming a skewed distribution of disturbances. Empirical examples are employed to illustrate the new methods developed herein and to discuss the merits and weaknesses in each of the two-step estimators.

Impact of Selection Bias from Mortality

For a baseline sample of I individuals and J follow-up time points, for convenience of analysis, a disability severity score, Y_{it}, is defined to indicate health status for individual i (i = 1, 2, …., I) at time t (t = 0, 1, …., J). It is then assumed that a hypothetical disability severity score exists instantaneously before dying for those who have been deceased between time (t - 1) and time t (t = 1, …., J). It is further assumed that the hypothetical disability severity score for the deceased, denoted by Y_{it}^{d}, is greater than or equal to a constant C_t, and the disability severity scores among survivors, Y_{it}^{s}, are all smaller than this constant.

Heckman's (1979) perspective serves to exhibit the impact of selection bias from mortality. Beginning with two longitudinal random-effects linear regression models, the complete model that includes all members of the baseline sample and a truncated model that consists of survivors only, given by

$$Y = X_1'\beta_1 + Z_1'\gamma_1 + \varepsilon_1 \qquad (1a)$$

$$Y \,|\, Y < C = X_2'\beta_2 + Z_2'\gamma_2 + \varepsilon_2, \qquad (1b)$$

where **Y** represents the (n × 1) vector of observed outcome data within the framework of a block design (n = I × [J + 1]). The matrix **X** is an (n × p) matrix for p − 1 independent variables and **Z** is a (n × r) design matrix for the random effects. The matrices **β** and **γ** are parameters for **X** and **Z** respectively. The random effects are assumed to be normally distributed with mean 0 and variance matrix **G**. The joint distribution of $\boldsymbol{\varepsilon_1}$ $\boldsymbol{\varepsilon_2}$ is assumed to be a singular distribution with covariance matrix $\boldsymbol{\sigma_{12}}$. While the residual term $\boldsymbol{\varepsilon_1}$ is assumed to be normally distributed with mean 0 and variance matrix $\boldsymbol{\sigma}_1^2$, it is implausible to assume that $\boldsymbol{\varepsilon_2}$ be normally distributed with zero

expectation, because the error term in (1b) may not be independent of the covariates.

Because $\mathbf{Y^d}$ is not observable, a dichotomous factor δ_{it} is defined to indicate the survival status for individual i between time (t – 1) and time t (t = 1, 2, …, J) and is used as a proxy for C, such that

$$\begin{cases} \delta_{it} = 0 \text{ if individual i dies between} \\ \quad \text{time (t-1) and t } \left(Y_{it} \geq C_t\right) \\ \delta_{it} = 1 \text{ if individual i survives from} \\ \quad \text{time } \left(\text{t-1}\right) \text{ and time t } \left(Y_{it} < C_t\right) \end{cases}.$$

Specifically, the disability severity score is viewed at time t as a joint distribution of two sequential events: the likelihood of survival between time (t - 1) and time t (S_t; t = 1,2, …., J) and the conditional density function on the disability severity score (Y_t) among those who have survived to t. Given the aforementioned assumptions, the expected disability severity score for individual i at time t can be estimated by the following equation

$$\mathrm{E}\left(Y_{it} \middle| X_{2i}, Z_{2i}, \delta_{it} = 1\right) = Pr\left(\delta_{it} = 1 \middle| X_{1i}\right) \left\{ \begin{array}{l} X'_{2i}\beta_2 + Z'_{2i}\gamma_2 \\ +E\left[\varepsilon_{2i} \middle| \varepsilon_{2i} < C_t - \left(X'_{1i}\beta_1 + Z'_{1i}\gamma_1\right)\right] \end{array} \right\}. \quad (2)$$

As demonstrated by (2), the conditional mean of the disturbance in the survivors sample is a function of $\mathbf{X}_{1i}$ and $\mathbf{Z}_{1i}$. The estimation of equation (2) without considering this correlation will lead to inconsistent parameter estimates and prediction biases. Therefore, modeling longitudinal processes of this disability severity score can be much beyond what a conventional single-equation linear regression can handle. Next, two refined longitudinal models are developed for reducing the selection bias in the analysis of longitudinal health data for older persons, one parametric and one nonparametric.

Parametric Joint Model

The parametric joint mixed model begins by constructing a selection model using survival rates as the dependent variable. Specifically, a Probit survival model is developed using the rationale of Heckman's (1979) two-step perspective to estimate the proportion surviving between time (t – 1) and time t (t =1, 2, …., J). Some empirical studies with joint modeling of longitudinal and survival data have used other statistical functions to estimate survival rates such as the Cox proportional hazard rate model and logistic regression (Egleston, Scharfstein, Freeman & West, 2006; Kurland & Heagerty, 2005; Leigh, Ward & Fries 1993; Pauler, McCoy & Moinpour, 2003; Ratcliffe, Guo & Ten Have, 2004). The Probit function is used here for convenience of illustration assuming survival probabilities are normally distributed. Specification of other functions would lead to the same results (Greene, 2003; Kalbfleisch & Prentice, 2002).

For individual i at time (t – 1), the probability of his or her survival to time t is given by

$$Pr\left(Y_{it} \middle| \delta_{it} = 1\right) = \Phi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right) \quad (3)$$
$$\text{t} = 1, 2, 3, ..., \text{J}$$

where Φ(.) represents the cumulative normal distribution function (Probit). From this equation, estimated survival rates can be obtained for each individual at J – 1 observation intervals. The estimates of $\Phi(\mathbf{X'\beta} + \mathbf{Z'\gamma})$ are then saved for each individual at each follow-up time point as an unbiased estimate of the survival rate.

Given the assumption that the hypothetical disability severity score for those who have been deceased between time (t – 1) and time t (t = 1, 2, …, J), the distribution of survivors' disability severity scores at time t is truncated on the right. Accordingly, the inverse Mills ratio for individual i at time t can be given by

$$\lambda_{it} = -\frac{\varphi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)}{\Phi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)} \text{ if } \delta_{it} = 1 \ (\text{Y} < \text{C}), \quad (4a)$$

$$\lambda_{it} = \frac{\varphi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)}{1-\Phi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)} \quad \text{if } \delta_{it} = 0 \ (Y \geq C), \tag{4b}$$

where φ(.) represents the standard normal density function. Values of λ's at time 0 (first wave) are all zero because no selection bias is present from deaths at the outset of the longitudinal investigation. As defined, the inverse Mills ratio for the deceased is the hazard rate of surviving between two adjacent time points; for those who have survived, it represents the risk of not surviving within an observational interval (Greene, 2003).

With the vector **λ** created, a conditionally unbiased truncated random-effects model is developed on the disability severity score at J time points, given by

$$Y\left(Y \middle| \delta = 1\right) = X'_2\beta_3 + Z'_2\gamma_3 + \sigma'_{e_2 v}\lambda + \varepsilon_3, \tag{5}$$

where $\boldsymbol{\sigma}_{ev}$ is a vector of covariance between $\boldsymbol{\varepsilon}_1$ and **v**, the latent error vector from (3), specified in the estimation process as a vector of the regression coefficients of **λ**, with elements assumed to be normally distributed. Because the survival rate and the disability severity score are inversely correlated, elements in $\boldsymbol{\sigma}_{ev}$ – with the exception of the first – are expected to take negative signs. With **λ** included in the estimation process, the error term $\boldsymbol{\varepsilon}_3$ is assumed to have mean 0 and variance σ_3^2, and to be uncorrelated with $\mathbf{X}_2$, $\mathbf{Z}_2$, and **λ**. When all assumptions on error distributions are satisfied, equation (5) generates unbiased and consistent parameter estimates because observations are presumably conditionally independent of each other.

Note that in equation (5), the inclusion of **λ** and **σ** accounts for the covariance between two error terms, $\boldsymbol{\varepsilon}_1$ and **v**, thereby indicating that the joint distribution of two sequential equations, represented by equation (2), is empirically embedded in (5).

Nonparametric Joint Random-Effects Model

The traditional two-step linear regression estimator and the joint longitudinal models depend on several strong assumptions regarding error distributional functions. When the assumption of multivariate normality for **ε** cannot be satisfied, as is often the case in health transitions (Liu, 2000; Manning, Duan & Rogers, 1987), Equation (5) cannot derive correct estimates for the underlying disability severity score. In these circumstances, Duan's (1983) and Liu's (2000) retransformation methods are extended into the context of repeated measures, assuming a nonparametric distribution of disturbances. One of the advantages of this approach is that researchers do not need to specify a parametric selection model to consider the missing data mechanisms. Rather, the selection bias is handled indirectly through estimating a smearing effect in the estimation process (Duan, 1983; Liu, 2000).

The log transformed nonzero value of the underlying disability severity score is used to address the possible non-linearity of its distribution among those with any disability. For this reason, a two-step procedure is proposed with the first equation meant to estimate the likelihood of having a nonzero disability score. The two-stage nonparametric mixed model is given by

$$\Pr\left(Y > 0\right) = \Phi\left(X'_2\beta_4 + Z'_2\gamma_4\right) \tag{6a}$$

$$log\left(Y \middle| Y > 0\right) = \left(X'_2\beta_5 + Z'_2\gamma_5 + \varepsilon_5\right)\xi, \tag{6b}$$

where ξ serves as a nonparametric adjustment factor for selection bias from high mortality. The expected disability severity score at various points in time can be expressed by the following joint distribution:

$$E\left(\hat{Y} \middle| S = 1\right) = \Phi\left(X'_2\hat{\beta}_4 + Z'_2\hat{\gamma}_4\right) exp\left(X'_2\hat{\beta}_5 + Z'_2\hat{\gamma}_5\right)\hat{\xi}. \tag{7}$$

As previously indicated, the distribution of the error term in health transition data is often skewed without following an identifiable pattern (Duan, 1983; Liu, 2000; Manning, Duan & Rogers, 1987). However, empirical data can be used to estimate values of **ξ** when the error distributional function is uncertain. First, assuming **X** to have full rank:

$$E(Y|Y>0) = E\left[log\left(X_2'\beta_5 + Z_2'\gamma_5 + \varepsilon_5\right)\right] = \int\left[log\left(X_2'\beta_5 + Z_2'\gamma_5 + \varepsilon_5\right)\right]dF(\varepsilon_5). \quad (8)$$

When the error distributional function F is unknown, this cumulative density function, F, is replaced by its empirical estimate $\hat{\mathbf{F}}_j$ at time-point t; this is referred to as the smearing estimate and is given by

$$E\left(\hat{Y}_t|Y_t>0\right) = E\int\left[log\left(X_{2t}'\beta_5 + Z_{2t}'\gamma_5 + \varepsilon_{5t}\right)d\hat{F}_{n_j}\left(\varepsilon_{5t}\right)\right] = \frac{1}{n_t}\sum_{i=1}^{n_t} log\left(X_{2it}'\beta_5 + Z_{2it}'\gamma_5 + \hat{\varepsilon}_{5it}\right) = log\left(X_{2t}'\hat{\beta}_5 + Z_{2t}'\hat{\gamma}_5\right)n_t^{-1}\sum_{i=1}^{n_{tj}} exp\left(\hat{\varepsilon}_{5it}\right), \quad (9)$$

where n_t is the number of observations at time t with nonzero disability severity scores and $\hat{\boldsymbol{\beta}}_5$ and $\hat{\boldsymbol{\gamma}}_5$ can be estimated by employing the maximum likelihood procedure without specifying a disturbance distributional function (Liu, 2000). When the sample size for a longitudinal study is large enough to derive a reliable expected value of errors, such a smearing estimate for the retransformation in log-linear equations is consistent, robust and efficient (Duan, 1983; Liu, 2000; Manning, Duan & Rogers, 1987).

The estimate of ξ at time t can be calculated by the equation

$$\xi_t = \frac{\sum_{i=1}^{n_t} exp\left[log\left(Y_{it}|Y_{it}>0\right) - \left(X_{2it}'\hat{\beta}_5 + Z_{2it}'\hat{\gamma}_5\right)\right]}{n_t}. \quad (10)$$

As presented, the nonparametric random-effects model does not depend on the specification of a given selection process; rather, it estimates an unknown error distribution by the empirical cumulative density function of the estimated regression residuals, and then takes the desired expectation with respect to the expected error distribution. If skeptical whether observations are conditionally independent, researchers might use the inverse Mills ratio as a covariate to account for the potential clustering among survivors thereby deriving more reliable parameter estimates. The complete dependence of this nonparametric approach on empirical data is obvious: If the longitudinal attrition due to reasons other than death is not random making the missingness non-ignorable, then the model-based predicted values of the disability severity score can be still severely biased.

Methodology

Illustrations

Data used for empirical demonstrations are from the Survey of Asset and Health Dynamics among the Oldest Old (AHEAD), a nationally representative investigation of older Americans. This survey, conducted by Institute of Social Research (ISR), University of Michigan, is funded by National Institute on Aging as a supplement to the Health and Retirement Study (HRS). At present, the survey consists of six waves of investigation; the Wave I survey was conducted between October 1993 and April 1994. Specifically, a sample of individuals aged 70 or older (born in 1923 or earlier) was identified throughout the HRS screening of an area probability sample of households in the nation. This procedure identified 9,473 households and 11,965 individuals in the target area range. AHEAD obtains detailed information on a number of domains, including demographics, health status, health care use, housing structure, disability, retirement plans and health and life insurance. Survival information throughout the six waves has been obtained by a link to the data of National Death Index (NDI). The present research uses data of all six waves: 1993, 1995, 1998, 2000, 2002 and 2004.

Disability severity, standing for an individual's health status in this study, is measured by a score of activities of daily living (ADL), instrumental activities of daily living (IADL), and other types of functional limitations (Liu, Engel, Kang & Cowan, 2005). A score of one is given to an individual who has any difficulty with a specific physical or social activity and the number of items for which difficulties are reported is then summed. As a

result, the score ranges from 0 (functional independence) to 15 (maximum disability). When predicting the survival rate (for the parametric joint model) or the probability of having any functional limitation (for the nonparametric joint model), such covariates as: veterans status (1 = veteran, 0 = non-veteran), age, gender (1 = female), education (years in school), ethnicity (1 = white, 0 = others), marital status (1 = currently married, 0 = other), smoking cigarettes and drinking alcohol, the number of serious health conditions, and self-rated health (5 scales: 1 = poor, 5 = excellent) are considered. The first four of these covariates (veteran status, age, gender and education) are used as the control variables in estimating the random-effects models and are rescaled to be centered about their means for analytic convenience. Specification of different sets of covariates at two different estimation stages helps reduce the occurrence of collinearity (Winship & Mare, 1992).

Three sets of the predicted number of functional limitations are compared at six time points; these are derived, respectively, from the conventional single-equation random-effects model, the parametric two-step joint model, and the nonparametric joint model. This provides the basis for examining how well each of these three random-effects longitudinal models fits the observed data for the following two reasons. First, if longitudinal dropouts due to reasons other than death are missing at random (MAR), the trajectory of the observed mean number of functional limitations is approximately unbiased. Here, the accurate description of empirical data serves as a criterion for the quality of a statistical model. Second, even if dropouts due to other reasons are missing not at random (MNAR), useful theoretical implications can be obtained by deviations of model-based predicted values from the empirical data.

The SAS PROC MIXED procedure with repeated measures is used to compute both fixed and random effects and to derive the predicted number of functional limitations at each time point (Littell, Milliken, Stroup, Wolfinger & Schabenberger 2006). Because intervals between two adjacent time points are unequally spaced in the AHEAD longitudinal data the REPEATED/TYPE = SP option was used in executing the SAS PROC.MIXED procedure to represent the autoregressive error structure of the data (Littell, et al., 2006). For analytic simplicity without loss of generality, between-individuals random effects are not further specified with the presence of a specific residual variance/covariance structure. Statistically, a combination of both error types is often found to fit the data about the same as does a model of either type (Hedeker & Gibbons, 2006). Hence, in the estimation process the variable time is treated as a series of dichotomous variables with the last time point, time 5 (time = 0, 1, 2, 3, 4, and 5), used as the reference.

Results

Table 1 presents the results of three random-effects models, the conventional, the parametric two-step and the nonparametric two-stage. In terms of the fixed effects, the intercept suggests the population estimate of the dependent variable at time 5 (year 2004); this time point is used as the reference in specification of five time dichotomous variables and all other covariates are centered about their sample means. The combined regression coefficients of the five time variables demonstrate an inverse-U shaped nonlinear function for the trajectory of transitions in the number of functional limitations, revealing the strong impact of the survival-of-the-fittest selection process among older Americans.

Of the control variables, veterans, older persons and women are expected to have a higher number of functional limitations than do their non-veteran, younger and male counterparts, other variable being equal. All regression coefficients, except those of veteran status, are statistically significant. The regression coefficient of lambda, the inverse Mills ratio, estimated for the parametric second-step random-effects model is sizable (-4.8184), statistically significant and takes a negative sign as expected. This suggests the importance of accounting for clustering effects when analyzing the longitudinal health data of older persons.

All estimates of the random effects are statistically significant. The SP variance/ covariance structure covers a relatively small but statistically significant portion of total variance for the conventional and the parametric two-step

random-effects longitudinal models. The relative size of this variance component increases considerably for the nonparametric random-effects model in which the dependent variable is the natural logarithm of the number of functional limitations among those with any functional limitation. The values of ξ's at the six time points, the adjustment factors in the means for the retransformation in the nonparametric random-effects model (not presented in Table 1) are, respectively, 1.3678 at time 0, 1.2448 at time 1, 1.1371 at time 2, 1.1491 at time 3, 1.1408 at time 4, and 1.2616 at time 5, all are statistically significant. The model Chi-square for each mixed model, reported in the last row of the table, is calculated as the difference in the value of -2 × (log likelihood) between the model with covariates and the model without any covariates.

Table 2 shows four sets of mean numbers of functional limitations in older Americans at six time points - 1993, 1995, 1998, 2000, 2002 and 2004 - derived from observed data and the three types of longitudinal random-effects models, respectively. Compared to the observed data, the conventional single-equation linear random-effects model systematically overestimates the number of functional limitations at every time point except the baseline and this overestimation increases as the survey progresses. The parametric two-step longitudinal joint model somewhat reduces such overestimation, but the adjustment appears very limited and deviations from the observed data are still considerable and systematic. By contrast, the nonparametric longitudinal joint model derives the closest set of the estimates to describe transitions in the number of functional limitations in older Americans.

Table 1: Results of Three Random-Effects 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:			
Intercept	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] Results of the second-step mixed model; [b] 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 Point	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.

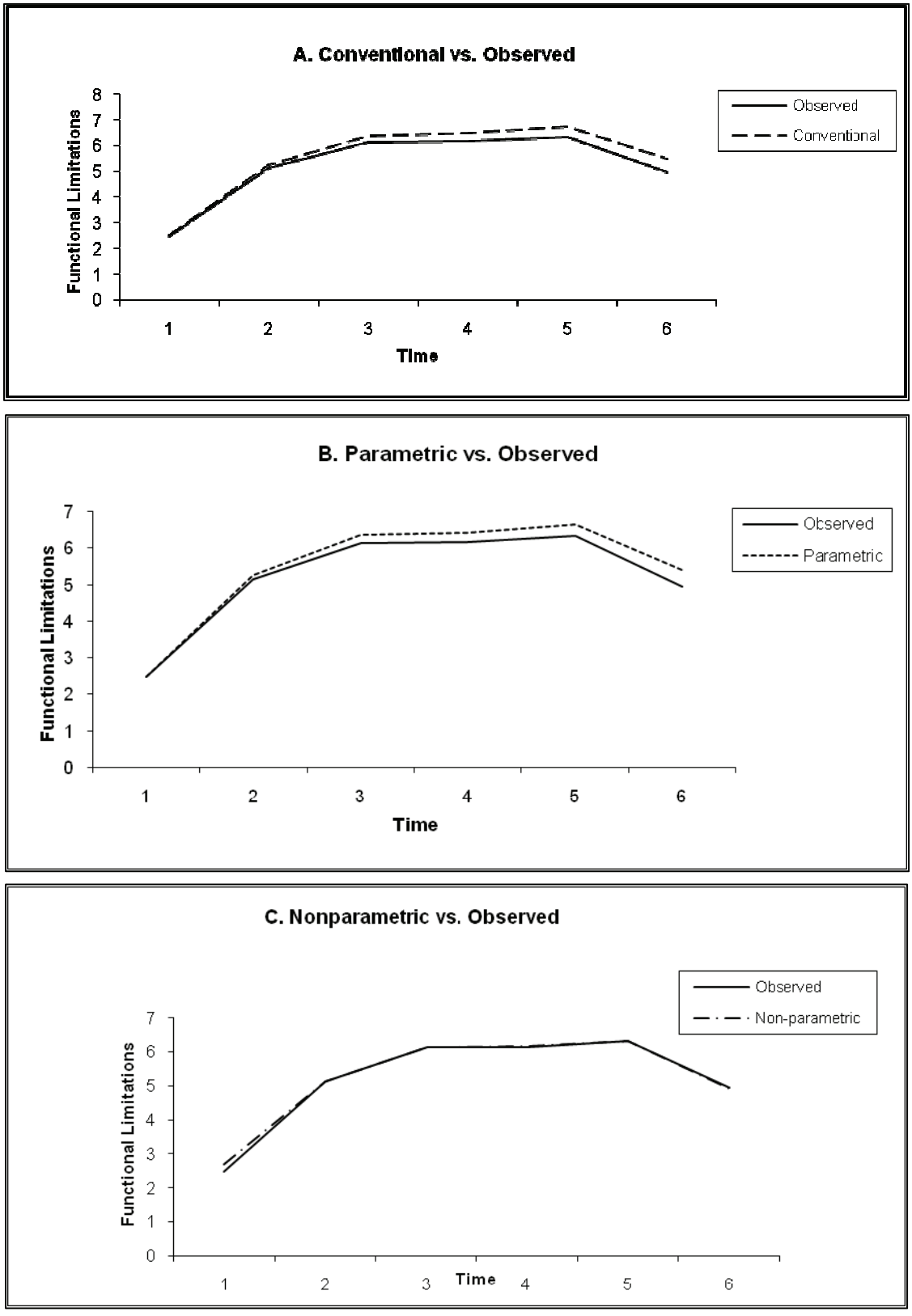
Figure 1 illustrates deviations in the predicted number of functional limitations derived from the three types of mixed models. Panel A compares the observed curve with the predicted values derived from the conventional single-equation random-effects model and shows distinct and systematic separations between the two growth curves. At each time point following the baseline survey, the predicted number of functional limitations obtained from the conventional single-equation random-effects model is considerably higher than the corresponding observed number. The predicted growth curve in Panel B, derived from the parametric longitudinal joint model, displays mitigated separation from the observed curve; however, the deviations remain sizable and systematic thereby reflecting the restriction of using parametric approach to correct for selection bias. In Panel C, the two curves almost coincide, demonstrating the accurate description of the empirical data by applying the nonparametric longitudinal joint modeling, which builds upon observed pattern of health transitions rather than impose strong assumptions on error distributions.

Conclusion

Non-ignorable missing data are important issues in longitudinal data analysis. Despite an abundant literature on this subject, none of the currently existing statistical models has the capacity to handle all types of non-ignorable dropouts (Hogan, Roy & Korkontzelou, 2004). Most models of this type are created for the analysis of longitudinal missing data in clinical experimental studies where repeated measures are often narrowly spaced and mortality is almost nonexistent. With respect to large-scale longitudinal data of older persons, currently available models are not specifically developed to reflect the unique influence of high mortality on estimating and predicting health outcomes at older ages. Because those who have been deceased between assessment periods no longer exist, various assumptions on the measurability of health status for dropouts are not plausible and meaningful.

When mortality rates are high, the direct application of conventional random-effects linear models on longitudinal health data can be associated with serious selection bias. As previously noted, mechanisms leading to biases on parameter estimates have been well documented (Egleston, Scharfstein, Freeman & West, 2006; Hogan, Roy & Korkontzelou, 2004; Kurland, & Heagerty, 2005; Leigh, Ward & Fries, 1993; Liu, 2000; Manning, Duan & Rogers, 1987; Pauler, McCoy & Moinpour, 2003; Ratcliffe, Guo & Ten Have, 2004). This study introduced two refined random-effects joint models and sought to substantially reduce

Figure 1: Transitions in Functional Limitations in Older Americans:
Growth Curves Derived from Three Approaches



bias incurred from changes in the distribution of health outcome data at multiple time points. The parametric longitudinal model is an extension of Heckman's (1979) traditional two-step estimator which, like other parametric joint models, is based on several restrictive assumptions on the joint modeling and error distributional functions. Researchers have questioned and discussed the validity and reliability of this type of two-step estimator. Much of the literature about this estimator focuses on the ill effects of violations against assumptions regarding λ, X and the error distributions (Demirtas, 2004; Fu, Winship & Mare, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little & Rubin 2003; Manning, Duan & Rogers, 1987; Winship & Mare, 1992).

This study shows that - as an extended case of the Heckman's perspective - the parametric two-step random-effects joint model has the capacity to reduce some of the deviations from the observed data; however, the degree of this adjustment is limited and deviations remain considerable and systematic. The limited effects of this approach are further evidenced by the similarity between the growth curve derived from this two-step estimator and the curve from the single-equation random-effects model (see Table 1 and Figure 1). In view of the difficulty in verifying assumptions on parametric distributional functions at multiple time points, the use of a nonparametric approach seems a more promising way of modeling longitudinal health data for older persons.

In reality, it is not possible to verify or contradict whether missingness is random by examination of the observed data (Demirtas, 2004; Little & Rubin, 2002). However, if non-death dropouts are missing at random, the selection bias from high mortality rates can be identified by examining the model fitness with observed health transition data. In many empirical applications in which mortality is low, the true cause of the missingness is often thought to be an unmeasured variable that is only moderately correlated with the response, not the response itself. Failure to account for the cause seems to introduce only minor bias (Schafer & Graham, 2002). If this phenomenon can be viewed as a general rule, the agreement of the model-based longitudinal trajectory with the observed curve can be used to measure the sensitivity of predicted health scores in older persons. The nonparametric longitudinal joint model presented herein is created particularly to correct for the selection bias from high mortality rates when the observed data are trustworthy and the non-death longitudinal dropouts are missing at random and thereby ignorable. This nonparametric regression model has the added advantage that the selection information (survival in the present study) does not need to be accounted for directly in the estimation process.

Because the nonparametric approach presented is meant to correct for the selection bias using empirical adjustments, its application must be based on researchers' confidence that biases from ignoring missing data from other causes are minor (Little, 1995). Therefore, its practicality is limited within the circumstances that non-response due to mortality is the only source of non-ignorable dropouts.

If non-death dropouts are missing not at random (MNAR), which is thought to be exceptional by some researchers (Schafer & Graham, 2002), investigators need to compare results generated from various statistical models handling non-ignorable dropouts, such as selection, semi-parametric, pattern-mixture models (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little, 1995; Pauler, McCoy & Moinpour, 2003; Robins, Rotnitzky & Zhao, 1995), and the present nonparametric joint approach. However, the effects of dropouts from different reasons on the longitudinal selection bias should be dealt with separately before a unified statistical model handling multi-cause dropouts can be eventually developed (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004). For example, dropouts due to mortality, sickness, migration or difficulty in answering sensitive questions may each involve a unique missing data mechanism. To fulfill this task, researchers must collect as much information as possible about various reasons for dropouts and incorporate this information into model development (Little, 1995).

Acknowledgement

This research was supported partly by the National Institute on Aging (NIH/NIA Grant No.: R03AG20140-01). Address correspondence to Dr. Xian Liu, Department of Psychiatry, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814. Email: Xian.Liu@usuhs.edu.

References

Hakan. D. (2004). Modeling incomplete longitudinal data. *Journal of Modern Applied Statistical Methods*, *3*, 305-321.

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, *78*, 605-610.

Egleston, B. L., Scharfstein, D. O., Freeman, E. E., & West, S. K. (2006). Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, *8*, 526-545.

Fu, V. K., Winship, C., & Mare, R. D. (2004). Sample selection bias models. In *Handbook of data analysis*, M. Hardy & A. Bryman (Eds.), 409-430. London: Sage.

Greene, W. H. (2003). *Econometric analysis* (*5th Ed.*). New Jersey: Prentice.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Annals of Econometrica*, *47*, 153-161.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Hogan, J. W., Roy, J., & Korkontzelou, C. (2004). Tutorial in biostatistics: handling drop-out in longitudinal studies. *Statistics in Medicine*, *23*, 1455-1497.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (*2nd Ed.*). New York: Wiley.

Kurland, B. F., & Heagerty, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by death. *Biostatistics*, *6*, 241-258.

Leigh, J. P., Ward, M. M., & Fries. J. F. (1993). Reducing attrition bias with an instrumental variable in a regression model: results from a panel of rheumatoid arthritis patients. *Statistics in Medicine*, *12*, 1005-1018.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (*2nd Ed.*). Gary, NC: SAS Institute, Inc.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (*2nd Ed.*). New York: Wiley.

Liu, X. (2000). Development of a structural hazard rate model in sociological research. *Sociological Methods & Research*, *29*, 77-117.

Liu, X., Engel, C. C., Kang, H., & Cowan, D. (2005). The Effect of Veteran Status on Mortality among Older Americans and its Pathways. *Population Research and Policy Review*, *24*, 573-592.

Manning, W., Duan, N., & Rogers, W. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics*, *35*, 59-82.

Pauler, D. K., McCoy, S., & Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine*, *22*, 795-809.

Ratcliffe, S. J., Wensheng G., & Ten Have, T. R. (2004). Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*, *60*, 892-899.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106-121.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Reviews of Sociology*, *18*, 327-350.

Yao, Q., Wei, L. J., & Hogan, J. W. (1998). Analysis of incomplete repeated measurements with dependent censoring times. *Biometrika*, *85*, 139-149.